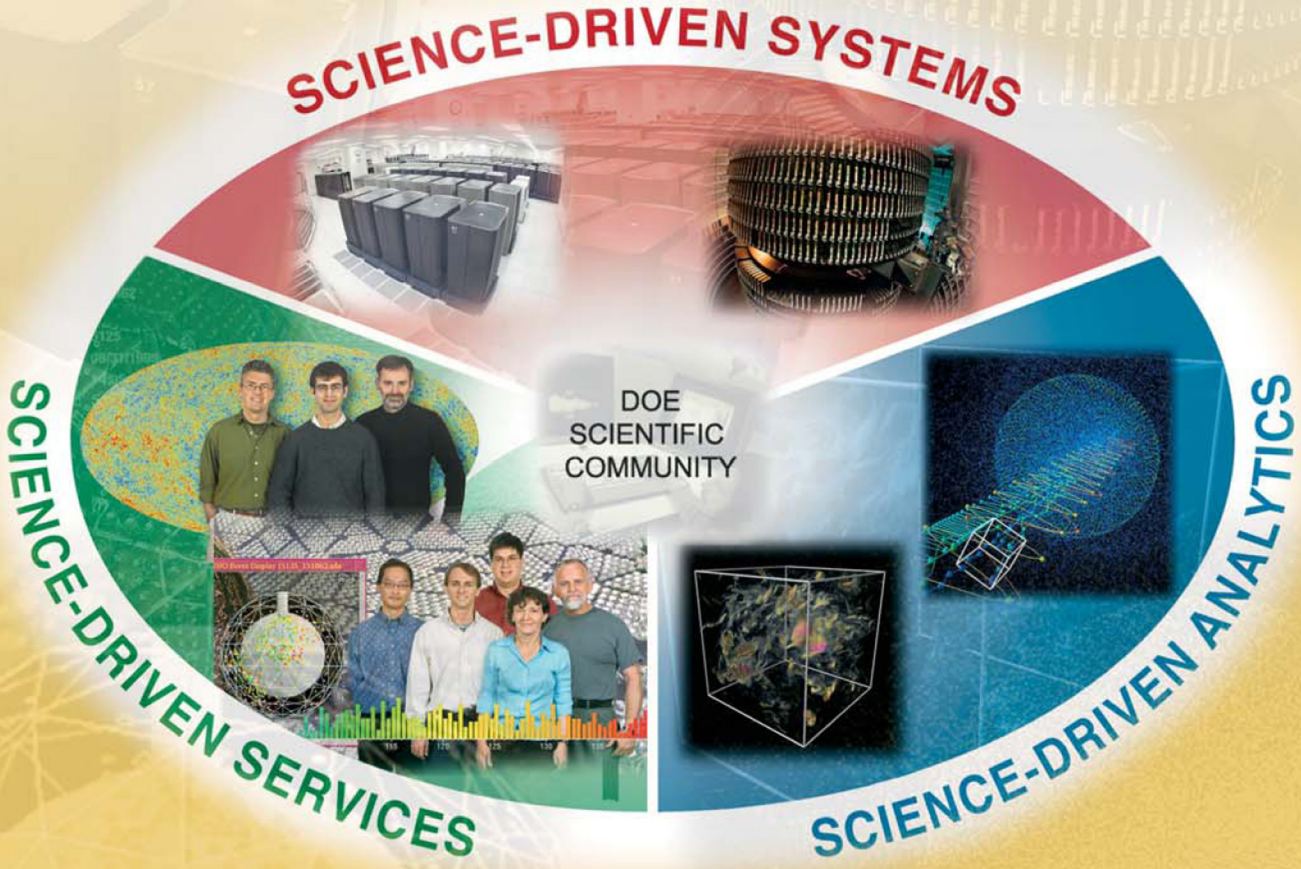
LBNL-57582

Science-Driven Computing: NERSC's Plan for 2006–2010

NERSC

NATIONAL ENERGY RESEARCH
SCIENTIFIC COMPUTING CENTER



BERKELEY LAB
ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Office of Science
U.S. DEPARTMENT OF ENERGY

LBNL-57582

Science-Driven Computing: NERSC's Plan for 2006–2010

Horst D. Simon, William T. C. Kramer, David H. Bailey, Michael J. Banda, E. Wes Bethel, Jonathon T. Carter, James M. Craw, William J. Fortney, John A. Hules, Nancy L. Meyer, Juan C. Meza, Esmond G. Ng, Lynn E. Rippe, William C. Saphir, Francesca Verdier, Howard A. Walter, Katherine A. Yelick

NERSC Center Division
Ernest Orlando Lawrence Berkeley National Laboratory
University of California
Berkeley, California 94720

May 2005

This work was supported by the Director, Office of Science, Office of Advanced Scientific Computing Research of the U.S. Department of Energy under Contract No. DE-AC 03-76SF00098.

Table of Contents

Executive Summary 1

1. Introduction: Science-Driven Computing 2
   1.1 NERSC's Mission 2
   1.2 A Science-Driven Strategy to Increase Scientific Productivity 2
      Science-Driven Systems 4
      Science-Driven Services 4
      Science-Driven Analytics 5
   1.3 A Key Resource for the DOE Office of Science 5

2. Science Requirements 9
   2.1 Nanoscience 11
   2.2 Climate Modeling 11
   2.3 Chemistry 12
   2.4 Fusion 13
   2.5 Combustion 13
   2.6 Astrophysics 14
   2.7 Biology 15
   2.8 Conclusion 15

3. Technology Challenges 2006–2010 16
   3.1 Processor Trends 2006–2010 17
   3.2 Interconnect Trends 2006–2010 18
   3.3 System Software and Tools 19

4. Science-Driven Systems 21
   4.1 NERSC's Response to Technology Challenges 21
   4.2 Computational Systems 22
   4.3 Computational Technology Choices 23
   4.4 Computational System Sizing 25
   4.5 Facility-Wide File System 25
   4.5 Mass Data Storage 26
   4.6 Networking 30
   4.7 Visualization and Analysis Systems 32
   4.8 Science-Driven Systems Summary 33

5. Science-Driven Services 36
   5.1 Direct Scientific Support 36
      Science-Driven Consulting 37
      Applications Software Support 37
      Web Information, Documentation, and Training 39
      Account Management and Allocations Support 39
   5.2 Collaborative Scientific Team Support 40
   5.3 Computing Infrastructure Support 41
      System and Network Monitoring and Support 41
      System Management 41
      System Improvements 42
      System Software Infrastructure 42

Security 43
Server Support 45
5.4 Outreach 45
5.5 Readiness for New Challenges 46

6. Science-Driven Analytics 46
6.1 The Role of Analytics in Scientific Research 47
6.2 NERSC's Analytics Strategy 48
Taking a Proactive Role in Deploying Emerging Technologies 49
Enhancing NERSC's Data Management Infrastructure 50
Expanding NERSC's Visualization and Analysis Capabilities 51
Enhancing NERSC's Distributed Computing Infrastructure 52
Understanding the Analytics Needs of the User Community 53

7. Investment Strategy and Budget 53
7.1 Staffing 54
7.2 Budget 54

8. Milestones 56

Executive Summary

NERSC traditionally provides systems and services that maximize the scientific productivity of its user community. NERSC takes pride in its reputation for the expertise of its staff and the high quality of services delivered to its users. To maintain its effectiveness, NERSC proactively addresses new challenges. Based on our interactions with the NERSC user community and our monitoring of technology trends, we observe three trends that NERSC needs to address over the next several years:

- the widening gap between application performance and peak performance of high-end computing systems
- the recent emergence of large, multidisciplinary computational science teams in the Department of Energy (DOE) research community
- the flood of scientific data from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows.

NERSC's responses to these trends are the three components of the science-driven strategy that NERSC will implement and realize in the next five years: ***science-driven systems, science-driven services,*** and ***science-driven analytics.*** This balanced set of objectives will be critical for the future of the enterprise and its ability to serve the DOE scientific community.

- ***Science-Driven Systems:*** Balanced introduction of the best new technologies for complete computational systems — computing, storage, networking, visualization and analysis — coupled with the activities necessary to engage vendors in addressing the DOE computational science requirements in their future roadmaps.
- ***Science-Driven Services:*** The entire range of support activities, from high-quality operations and user services to direct scientific support, that enable a broad range of scientists to effectively use NERSC systems in their research. NERSC will concentrate on resources needed to realize the promise of the new highly scalable architectures for scientific discovery in multidisciplinary computational science projects.
- ***Science-Driven Analytics:*** The architectural and systems enhancements and services required to integrate NERSC's powerful computational and storage resources to provide scientists with new tools to effectively manipulate, visualize, and analyze the huge data sets derived from simulations and experiments.

NERSC is ready to respond to significant changes that are expected in high performance computing (HPC) technologies. Obtaining high performance for scientific applications will require NERSC's active involvement, in partnership with the larger HPC community, in understanding, driving, and adopting these technologies. With its experienced and innovative staff, NERSC is well poised to meet these challenges.

1. Introduction: Science-Driven Computing

1.1 NERSC's Mission

The mission of the National Energy Research Scientific Computing Center (NERSC) is to accelerate the pace of scientific discovery by providing high performance computing, information, data, and communications services for research sponsored by the DOE Office of Science (SC). NERSC is the principal provider of high performance computing services for the capability needs of Office of Science programs — Fusion Energy Sciences, High Energy Physics, Nuclear Physics, Basic Energy Sciences, Biological and Environmental Research, and Advanced Scientific Computing Research.

Computing is a tool as vital as experimentation and theory in solving the scientific challenges of the twenty-first century. Fundamental to the mission of NERSC is enabling computational science of scale, in which large, interdisciplinary teams of scientists attack fundamental problems in science and engineering that require massive calculations and have broad scientific and economic impacts. Examples of these problems include global climate modeling, combustion modeling, magnetic fusion, astrophysics, computational biology, and many more. NERSC uses the Greenbook process (described in Section 2) to collect user requirements and drive its future development.

Lawrence Berkeley National Laboratory (Berkeley Lab) operates and has stewardship responsibility for NERSC, which, as a national resource, serves about 2,400 scientists annually throughout the United States. These researchers work at DOE laboratories, other Federal agencies, and universities (over 50% of the users are from universities). Computational science conducted at NERSC covers the entire range of scientific disciplines, but is focused on research that supports DOE's missions and scientific goals.

1.2 A Science-Driven Strategy to Increase Scientific Productivity

Since its founding in 1974, NERSC has provided systems and services that maximize the scientific productivity of its user community. NERSC takes pride in its reputation for the expertise of its employees and the high quality of services delivered to its users. To maintain its effectiveness, NERSC proactively addresses new challenges. Based on our interactions with the NERSC user community and our monitoring of technology trends, we observe three trends that NERSC needs to address over the next several years:

- the widening gap between application performance and peak performance of high-end computing systems
- the recent emergence of large, multidisciplinary computational science teams in the DOE research community
- the flood of scientific data from both simulations and experiments, and the convergence of computational simulation with experimental data collection and analysis in complex workflows.

NERSC's responses to these trends are the three components of the science-driven strategy that NERSC will implement and realize in the next five years: ***science-driven systems, science-driven services,*** and ***science-driven analytics*** (Figure 1). This balanced set of objectives will be critical for the future of the enterprise and its ability to serve the DOE scientific community.

- ***Science-Driven Systems:*** Balanced introduction of the best new technology for complete computational systems — computing, storage, networking, visualization and analysis — coupled with the activities necessary to engage vendors in addressing the DOE computational science requirements in their future roadmaps.
- ***Science-Driven Services:*** The entire range of support activities, from high-quality operations and user services to direct scientific support, that enable a broad range of scientists to effectively use NERSC systems in their research. NERSC will concentrate on resources needed to realize the promise of the new highly scalable architectures for scientific discovery in multidisciplinary computational science projects.
- ***Science-Driven Analytics:*** The architectural and systems enhancements and services required to integrate NERSC's powerful computational and storage resources to provide scientists with new tools to effectively manipulate, visualize, and analyze the huge data sets derived from both simulation and experiment.

Figure 1. Conceptual diagram of NERSC's plan for 2006–2010.

Science-Driven Systems

Applications scientists have been frustrated by a trend of stagnating application performance relative to dramatic increases in claimed peak performance of high performance computing systems. This trend has been widely attributed to the use of commodity components whose architectural designs are unbalanced and inefficient for large-scale scientific computations. It was assumed that the ever-increasing gap between theoretical peak and sustained performance was unavoidable. However, results from the Earth Simulator (ES) in Japan clearly demonstrate that a close collaboration with a vendor to develop a science-driven solution can produce a system that achieves a significant fraction of peak performance for critical scientific applications. The key to the ES success was the long-term collaborative development strategy between the scientists of JAMSTEC (Japan Marine Science and Technology Center) and NEC Corporation.

Realizing that effective large-scale system performance cannot be achieved without a sustained focus on application-specific systems development, NERSC has begun a science-driven systems strategy. The goal of this effort is to influence the vendors' product roadmaps to improve system balance and to add key features that address the requirements of demanding capability applications at NERSC — ultimately leading to a sustained Pflop/s system for scientific discovery. This strategy involves extensive interactions between domain scientists, mathematicians, computer experts, as well as leading members of the vendors' research and product development teams. These interactions will be pursued in collaboration with other DOE laboratories, in particular with future advanced computing research testbed (ACRT) activities funded by SC.

As Section 3 will explain, NERSC must be prepared for disruptive changes in processor, interconnect, and software technologies. Obtaining high application performance will require the active involvement of NERSC in understanding, driving, and adopting these technologies. The move towards open-source software will require additional efforts in software integration at NERSC.

The goal of the science-driven systems strategy is to enable new scientific discoveries, and that requires a high level of sustained system performance on scientific applications. The NERSC approach takes into account both credibility and risk in evaluating systems and will strike a balance between innovation and performance on the one hand and reliability on the other. While the discussion often focuses on the high-end platforms, NERSC will continue to emphasize maintaining Center balance, that is, improving all the systems at NERSC — storage, networking, visualization and analysis — commensurately with improvements in the high-performance computing platforms.

Science-Driven Services

The DOE computational science community, in all its disciplines, has been organizing itself into large multidisciplinary teams. This trend was driven by the DOE Scientific Discovery through Advanced Computing (SciDAC) initiative, but has reached beyond the SciDAC teams. This trend away from the single principal investigator model for computational science, which has dominated computing in the natural sciences for most of the second half of the last century, has been driven by necessity as well as opportunity. The transformation became most apparent after massively parallel computers came to dominate the high end of available computing resources.

The gap between the peak performance of today's terascale platforms and the performance attained by simply porting an application to run on them increased dramatically in comparison with the typical performance of scientific codes on the vector supercomputers of the 1980s and early 1990s.

NERSC has been focused on working with computational scientists to close this gap and help them scale their applications efficiently to current platforms. Technology trends that will be described in more detail in Section 3 indicate that the gap between the peak performance of next-generation systems and performance that is easily attainable could increase even more. NERSC has formulated a science-driven services strategy that will address the requirements of these large computational science teams even more so than in the past, while at the same time maintaining the high level of support for all of its users.

Science-Driven Analytics

A major trend occurring in computational science and in particular in the portfolio of the Office of Science is the flood of scientific data from both simulations and experiments, and the convergence of experimental data collection, computational simulation, visualization, and analysis in complex workflows. Deriving scientific understanding from massive datasets is a growing challenge. The fact that the Office of Science's mission includes building and operating major experimental facilities — including the light sources and neutron sources used by the nation's chemists, materials scientists, and biologists — makes managing and analyzing the data from these experiments, as well as simulations, a key part of NERSC's mission as well.

In recent years NERSC has seen a dramatic increase in the data arriving from DOE-funded research projects. This data is stored at NERSC because NERSC provides a reliable long-term storage environment that assures the availability and accessibility of data for the community. NERSC has helped accelerate this development by deploying Grid technology on all of its systems and by enabling and tuning high performance wide area network connections to major facilities, for example the Relativistic Heavy Ion Collider at Brookhaven National Laboratory.

Now NERSC must invest resources to complete an environment that allows easier analysis and visualization of large datasets derived from both simulation and experiment. Our third new thrust in science-driven analytics will enable scientists to combine experiment, simulation, and analysis in a coordinated workflow. This thrust will include activities enhancing NERSC's data management infrastructure, expanding NERSC's visualization and analysis capabilities, enhancing NERSC's distributed computing infrastructure, and understanding the analytics needs of the user community (see Section 6).

1.3 A Key Resource for the DOE Office of Science

In "Facilities for the Future of Science: A Twenty Year Outlook," the Office of Science has identified the need for creating new and/or improving on the current computational capability as a critical aspect of realizing its advanced scientific computing research vision.[1] It identified the

[1] *Facilities for the Future of Science: A Twenty-Year Outlook* (Washington, DC: U.S. Department of Energy, Office of Science, November 2003).

NERSC upgrade as a near-term priority to ensure that NERSC, DOE's premier scientific computing facility for unclassified mission-critical research, continues to provide high-performance computing resources to support the requirements of scientific discovery.

As a high-end facility that serves all the DOE SC programs with capability and high-end capacity resources, NERSC is a key resource in SC's portfolio of computing facilities. NERSC has established a reputation for providing reliable and robust services along with unmatched support to its users. Because of investments such as SciDAC, and the important role that computation will play in Genomics:GTL (formerly Genomes to Life) and the Nanoscale Science Research Centers, demands for computational resources in SC will continue to grow at a rapid rate, and NERSC's growth must keep pace. The new National Leadership Computing Facility (NLCF) at Oak Ridge National Laboratory will not close this resource gap, because the NLCF will support a small number (10–20) of projects that require the highest computational capability for extended periods of time. The NLCF is a national resource for computational science whose mission spans multiple agencies. In contrast, NERSC supports a large number (200–300) of projects of medium to large scale, occasionally needing a very high capability resource, that fall within the mission of the Office of Science. The scientific productivity enabled by NERSC is demonstrated by the 2,206 papers in refereed publications in 2003 and 2004 that were based at least in part on work done at NERSC.

In NERSC's experience there is a continuum of scientific computing systems and facilities, as represented in Figures 2 and 3. There are a few research groups with experienced users and very high computational requirements who are in a good competitive position to use the NLCF. There is a much larger number of PIs and projects with high-end requirements who are best served by NERSC's high-end systems and comprehensive services, both of which distinguish NERSC from leadership computing and midrange computing centers, such as institutional or departmental clusters. Capability users include both single-principal-investigator teams and community science teams. NERSC's science-driven services are important for both types of high-end users.

NERSC supports large-scale teams working on advanced modeling and simulation "community codes" whose development is shared by entire scientific research communities. These codes employ new mathematical models and computational methods designed to better represent the complexity of physical process and to take full advantage of current computational systems. NERSC provides focused support for these teams, with the goal of bridging the software gap between currently achievable and peak performance on terascale platforms, as was explicitly stated in the SciDAC plan.

NERSC also supports single-PI teams consisting of a lead researcher and his or her group of collaborators, postdocs, and students, usually concentrated at a single location, working on a research-level code which is not shared outside the collaboration, or using well established third-party applications software. For this class of users, NERSC's science-driven service is important because they usually are less knowledgeable about computational technologies and they lack the resources to establish in-depth collaborations with computer science or mathematics experts. NERSC's services fill this gap and assure an efficient utilization of the computational resources. NERSC helps single-PI projects to make the transition from their desktop/cluster environment to

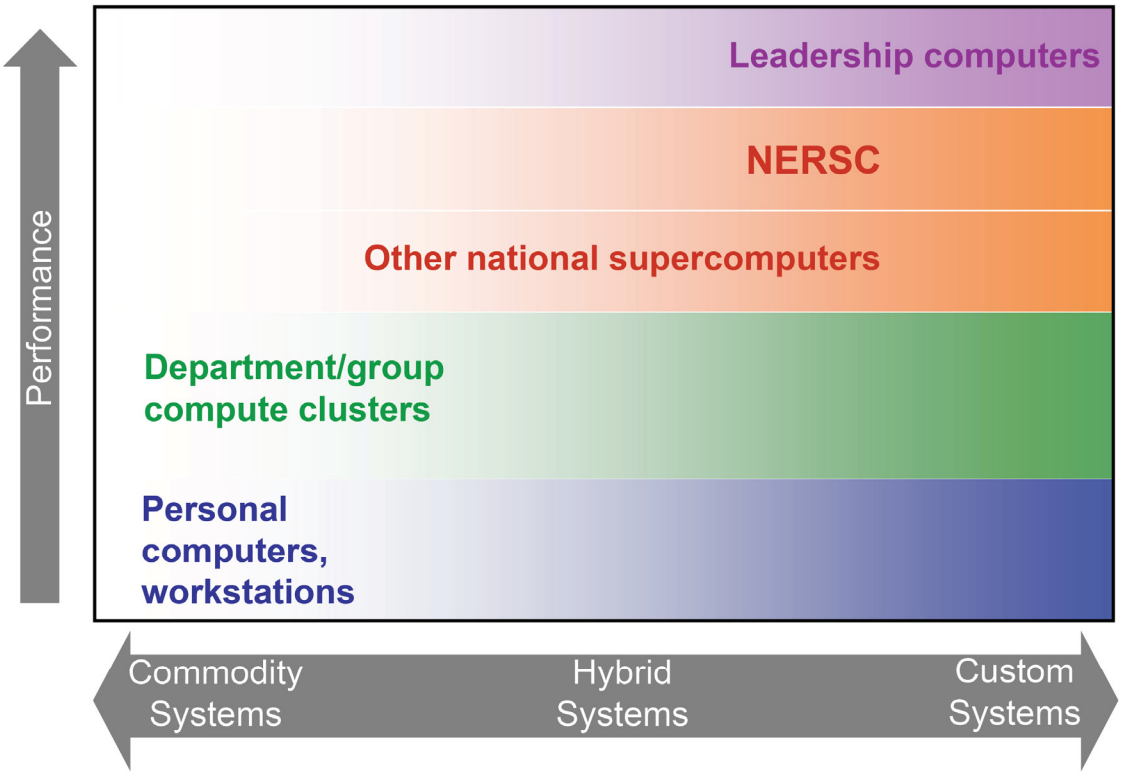


Figure 2. The continuum of scientific computing systems.

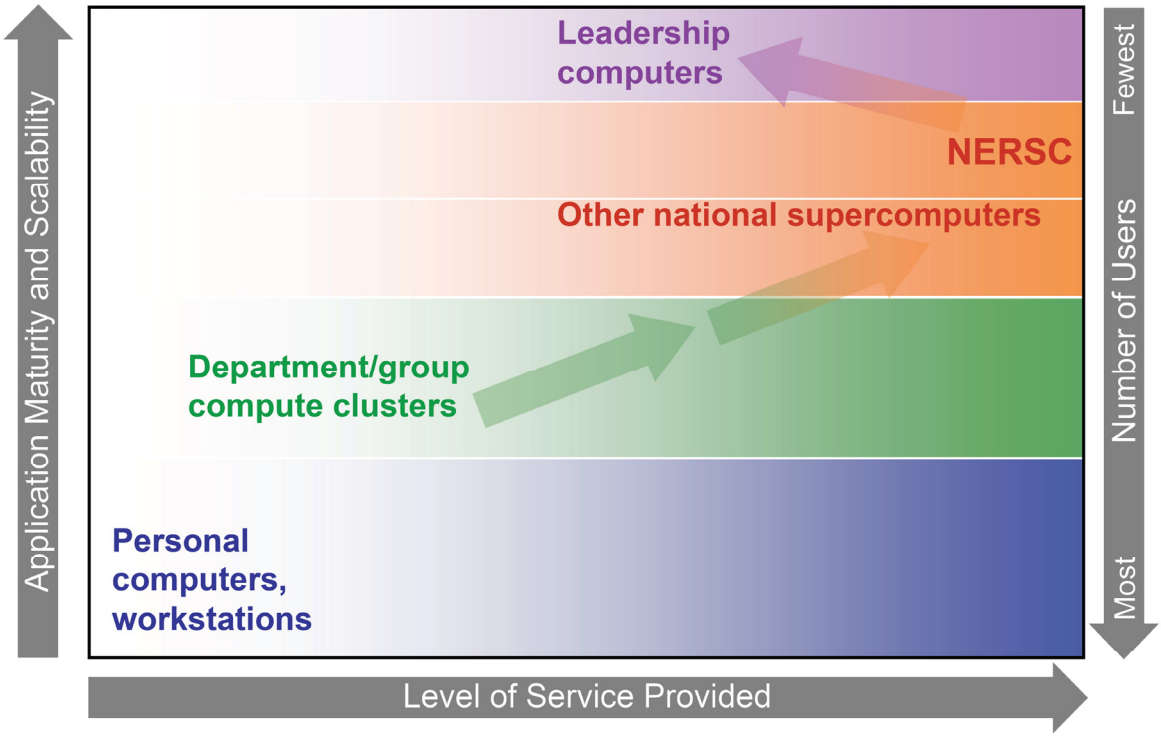


Figure 3. The continuum of scientific computing facilities.

the high end at NERSC, and enables them to compute with orders of magnitude more parallelism and higher efficiency.

Computing at NERSC not only produces important scientific insights but also gives these users and teams the opportunity to advance to the leadership computing level for their most challenging computations. NERSC will closely collaborate with the NLCF to facilitate the transition of computational science projects between these two facilities.

NERSC, as a centralized facility, properly staffed and managed, provides the best possible mechanism for technology transfer between the computational efforts of different research programs. Moreover, a concentration of computing resources provides a more flexible mechanism to address changing priorities. SC's priorities for its programs sometimes change quickly because it is a mission agency. A general-purpose facility like NERSC, with a staff prepared to support the broadest possible array of scientific disciplines, allows DOE to switch priorities and quickly apply its most powerful computing resources to new challenges.

NERSC's role as a general scientific computing facility requires it to provide resources that are of common utility to the programs of the Office of Science. However, NERSC must be responsive to the specific needs of each program. Specific support for different programs, tailored to their varying needs, has been a key to the success of the Center. Examples range from the collaborative effort of NERSC staff in scaling INCITE applications to 2,048 and 4,096 processors, to the operation of the PDSF cluster for the high energy and nuclear physics communities. The breadth of NERSC's support is best expressed by Figures 4 and 5, which summarize NERSC usage by discipline and institution. It is important to note that until recently NERSC was the only computational facility in the U.S. to provide access and support for both academic and research laboratory users.

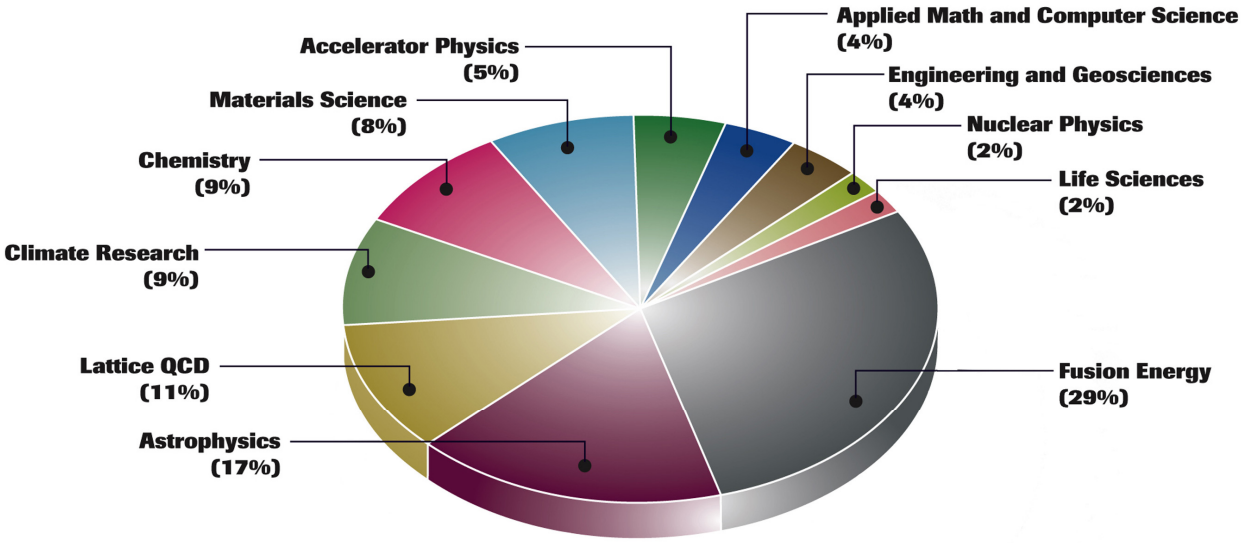


Figure 4. NERSC usage by scientific discipline for FY2004.

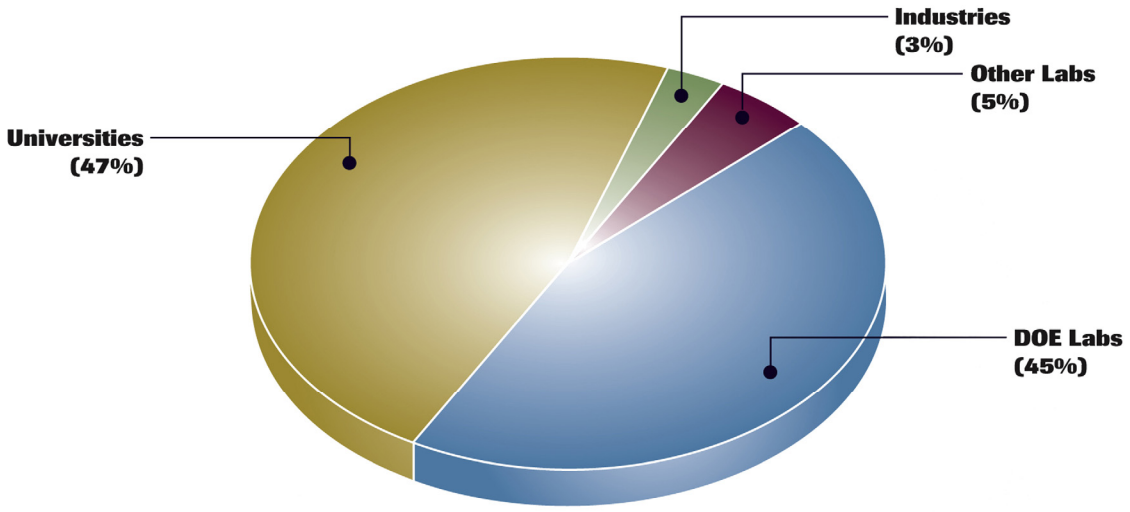


Figure 5. NERSC users by institution type for FY2004.

2. Science Requirements

Science requirements for NERSC during the next five years are expected to dramatically increase over current levels for a variety of reasons. Well established fields of computational science, such as fusion and climate modeling, are moving to more physically sophisticated and higher-resolution models. Relative newcomers to high-end computing, especially the biological sciences, are seeing rapidly accelerating growth in the importance of computation in their research. And data-intensive simulations and experiments are requiring new and improved capabilities for data management and analysis.

For these reasons, we see aggregate computational requirements increasing faster than Moore's Law through the end of the decade. In addition, we foresee a shift in requirements, as data-intensive computations grow in both size and number, pointing to a greater need for large-scale dataset management, together with a greater need for visualization and analysis tools.

We present here an outline of these future requirements, categorized by sponsoring office:

- Office of Basic Energy Sciences: materials science (including nanoscience), chemical sciences (including combustion), geosciences, and energy biosciences.
- Office of Biological and Environmental Research: protein folding, Genomics:GTL, and global climate change.
- Office of Fusion Energy Sciences: plasma turbulence and transport, macroscopic stability, stellarator physics, electromagnetic wave/plasma interactions, intense ion beams and simulation of fast ignition.
- Office of High Energy Physics: lattice QCD, accelerator design, experimental particle physics, astrophysics.
- Office of Nuclear Physics: lattice QCD, nuclear structure, astrophysics, experimental nuclear physics.

- Office of Advanced Scientific Computing Research: SciDAC program, including both scientific applications projects and Integrated Software Infrastructure Centers (ISICs); ongoing non-SciDAC projects (ACTS Toolkit, etc.).

Space does not permit us to provide details on all of the above applications, so we will focus on a few key strategic applications. The algorithmic requirements of these applications are summarized in Table 1.

Table 1
Applications and Algorithms Matrix

Science Areas	Multi-physics & multi-scale	Dense linear algebra	Sparse linear algebra	FFTs	AMR	Data intensive
Nanoscience	X	X	X	X		
Climate	X			X	X	
Chemistry	X	X	X	X		
Fusion	X	X	X		X	X
Combustion	X		X		X	X
Astrophysics	X	X	X	X	X	X
Biology	X	X				X
Nuclear		X	X			X

From these algorithmic requirements, we can draw some general requirements for NERSC's systems:

- Multiphysics, multiscale algorithms run best on a general-purpose, balanced architecture because they use a variety of algorithms and techniques.
- Dense linear algebra requires a high-speed CPU and a high flop/s rate.
- Sparse linear algebra requires a high performance memory system.
- Multidimensional FFTs require a high interconnect bisection bandwidth.
- Adaptive mesh refinement needs good performance for irregular data and control flow. Complex and changing load balances pose special challenges to the scalability of AMR codes, favoring a smaller number of high performance nodes over a larger number of lower performance nodes.
- Data or I/O intensive computation requires advanced networking, mass storage, and data management capabilities.

We make a few general observations about the challenges facing parallel scientific applications. First, all areas of science today require a broad range of algorithms. Second, as the science has become more sophisticated, the applications and algorithms have become more complex — dense linear algebra, a common feature in scientific codes 20 years ago, is being supplanted by more efficient algorithms based on sparse linear algebra techniques, adaptive mesh refinement,

and other algorithms that are more difficult to parallelize and more challenging to processor and memory systems.

In this section we have identified the computational challenges facing scientific applications. NERSC has identified analytics challenges as another major factor in scientific productivity. These are described in detail in Section 6.

2.1 NANOSCIENCE

The fabrication and integration of nanoscale systems promises to revolutionize science and technology, from targeted drug delivery in medicine to ultra-fast single electron devices for computer technology. Ray Orbach, the Director of the DOE Office of Science, has identified nanoscale science as one of the seven highest priorities for the Office of Science. Computational simulations play an indispensable role in nanoscience research, in part because the complexities of nanosystems often make traditional analytical tools inapplicable, and the small size scales make direct experimental measurements very difficult.

Many-body methods in computational nanoscience are based either on a Monte Carlo approach (quantum Monte Carlo) or eigenfunction-type calculations, which involves the diagonalization of large matrices using dense or iterative solvers. In the single particle method, the wavefunctions are usually expanded in plane waves (Fourier components), and calculations typically involve parallel 3D fast Fourier transforms (FFTs) and dense linear algebra in the form of iterative eigensolvers. A third class of computation is classical molecular dynamics codes, which are used to study the synthesis of nanostructures and large nanostructures beyond the scope of quantum calculations.

Two specific applications that scientists in this area foresee include the following:

- **Electronic structures and magnetic materials.** The current state of the art is a 500-atom structure, which requires 1.0 Tflop/s sustained performance (typically for several hours), and 100 GB main memory. Future requirements (a hard disk drive simulation, for instance) are for a 5000-atom structure, which will require roughly 30 Tflop/s and 4 TB main memory.
- **Molecular dynamics calculations.** The current state-of-the-art is for a 10^9-atom structure, which requires 1 Tflop/s sustained and 50 GB memory. Future requirements (an alloy microstructure analysis) will require 20 Tflop/s sustained and 5 TB main memory.

2.2 CLIMATE MODELING

One of the nation's leading challenges is to understand the phenomenon of global warming, and to better assess future trends and possibilities for mitigation. An important short-term national objective is to increase the United States' contribution to the assessment reports of the Intergovernmental Panel on Climate Change. Current state-of-the-art climate models represent a complex coupling of models of the major climate subsystems, including atmosphere, ocean, sea ice, and land systems. Individual simulations are lengthy, often requiring many months to complete, and ensembles of simulations are required to gain confidence in the results.

Climate codes solve equations of hydrodynamics, radiation transfer, thermodynamics, and chemical reaction rates. Current approaches are marked by finite difference calculations acting on fairly regular spatial grids, requiring high main memory bandwidth. FFTs of a short to medium length are also used in current models, although these may be replaced in the future. Scalability of individual simulations tends to be poor relative to other advanced scientific applications because of modest concurrency in the codes (limited in part by the need for century-long simulations), although the ensemble requirement provides an embarrassingly parallel dimension to increase scientific throughput. It should be noted that existing sea ice and land codes are difficult to vectorize.

The current state of the art is roughly 1.4 degree spacing with 26 vertical layers for the atmosphere, and 1 degree spacing with 40 vertical layers in the ocean. Such simulations require roughly 140 seconds run time on 208 CPUs of Seaborg per simulated day. Near-term enhancements to code physics, including the introduction of carbon cycles, nitrogen cycles, biogeochemistry, and dynamic vegetation, will increase the computational requirements 25x with no increase in resolution. Planned near-term enhancements in resolution are 0.7 degree for the atmosphere and 1 degree for the ocean, resulting in a 5.75x increase in the cost of a coupled run; combined with the physics increases, this would result in 144x growth over the next three to four years. Proposed resolution enhancements in the five- to six-year time frame are a 0.35 degree atmosphere coupled with a 0.5 degree ocean, requiring 800–1000x current usage. At that resolution, achieving a throughput rate of five simulated years per compute day would require 898 Tflop/s sustained.

2.3 CHEMISTRY

DOE's Chemical Sciences program supports a major portion of the nation's research activities in computational chemistry. The program currently faces several challenges that will require orders-of-magnitude increase in computational capabilities, even assuming algorithmic improvements. Some of these challenges include crossing multiple length and time scales, predicting and controlling chemical reactivity, and designing chemical systems with desired properties.

Some applications that are beyond current capabilities, but could be addressed with a substantial increase in computational power include:

- Highly accurate thermodynamics and kinetics for the examination of reactions in large, realistic systems. The results of these calculations could be applicable in areas such as combustion and climate modeling.
- Calculations involving relativistic effects, including light element systems where extreme accuracy is needed, and even certain heavy element systems. Structures and ground state energetics are important, but also properties such as infrared frequencies, nuclear magnetic resonance shifts, and optical properties.
- Direct numerical simulation of flames, including modules to predict radiative transfer and condensed-state physics (soot particles, liquid fuel droplets, etc.).

2.4 FUSION

Fusion energy research is a crucial part of the DOE's long-term mission for energy independence. In particular "ITER for fusion energy" has been identified as the number one near-term priority for the Office of Science. Computational simulation supports experimental fusion research not only in the design of experimental facilities, but also in analysis of the resulting data and the development and validation of theory.

Global stability of the plasma confinement is an essential ingredient of a working reactor. Simulating the onset and evolution of instabilities and predicting confinement failure is particularly difficult because of the large range of time and length scales and high anisotropy of the plasma. Another important aspect of fusion energy is the balance between the heat generation of the burning plasma and the heat loss from electromagnetic turbulence. This behavior impacts the design and cost of the reactor, as it affects both material strength and cooling. Three other phenomena under study are the breaking and reconnection of oppositely directed magnetic field lines in a plasma, the mechanism for injecting fuel into the reactor, and edge plasma models. All of these must be understood much better before fusion energy can be a practical reality.

Each of these computations is a multi-physics, multi-scale computation involving numerous algorithms and computational techniques. Both regular and irregular access computations are involved. Two of the key numerical techniques employed are adaptive mesh refinement and advanced nonlinear solvers (particularly for the "stiff" partial differential equations that are often involved). Two sample applications include:

- A Tokamak simulation, namely an ion temperature gradient turbulence simulation in an ignition experiment. The current state of the art is a $3000 \times 1000 \times 64$ grid, or 2×10^8 grid points. Each grid cell contains eight particles, for a total of 1.6×10^9 particles; 50,000 time steps are required. The total operation count for such a calculation is 3×1017 flops, or 8 hours on a 10 Tflop/s system sustained, with 1.6 TB memory. Improved plasma models will increase run times by a factor of 10.
- The all-orders spectral algorithm (AORSA) code, which is used to study the effects of radio-frequency electromagnetic waves in plasmas. Such calculations currently involve a $120{,}000 \times 120{,}000$ complex linear system, which requires 1.3 hours on a 1 Tflop/s sustained system, with 230 GB memory. A $300{,}000 \times 300{,}000$ run would require eight hours. Future plans are for a $6{,}000{,}000 \times 6{,}000{,}000$ system, which will require 160 hours on a 1 Pflop/s sustained system, with 576 TB memory.

2.5 COMBUSTION

Combustion provides more than 85% of the nation's energy. Meeting U.S. energy demands and environmental concerns requires that power generation, industrial process, and transportation systems operate with higher efficiency and lower emissions. To that end, large-scale computational simulations are being employed to study the highly complicated turbulence-chemistry effects. These computations typically span a huge range in time and length scales (as much as a factor of 10^9), requiring large and adaptively managed sets of grid points.

Combustion codes typically combine a subset of the following computational techniques:

- explicit finite difference, finite volume, and finite element methods for systems of nonlinear partial differential equations (PDEs)
- implicit finite difference, finite volume, and finite element methods for elliptic and parabolic PDEs (typically utilizing iterative sparse linear solvers)
- zero-dimensional physics (often heterogeneous in work per point), including evaluation of thermodynamic and transport data as well as integration of stiff systems of ordinary differential equations
- adaptive mesh refinement
- Lagrangian particle methods embedded in finite difference, finite volume, and finite element algorithms.

Because of the huge dynamic range of time and length scales, it is very easy to construct important problems that are several orders of magnitude beyond current capabilities — three orders beyond current capability is very useful. The challenge of present-day researchers is to find algorithms that render certain approximate solutions feasible in reasonable run times.

2.6 ASTROPHYSICS

Another of the top priorities identified recently for the Office of Science is the question of "dark energy and the search for the genesis," namely the quest to understand some of the most profound scientific questions: What is the origin of all matter and energy? What is the fate of the Universe? What is the nature of space and of time? The recent discovery of dark matter, dark energy, and the accelerating Universe is but one example. With new observational tools, astrophysics is undergoing a transition from a data-starved to a data-swamped discipline, and has turned to computational simulations and data processing.

Important applications in the astrophysics arena include supernova hydrodynamics and energy transport, black hole simulations, and early universe field theory. These computations involve multi-physics (including thermodynamics, turbulence, nuclear chemistry, and others) and multi-scale phenomena, with a dynamic range in both time and space that can be enormous, requiring adaptive mesh refinement methods. Key underlying algorithms include dense linear algebra, FFTs, spherical harmonic transforms, and operator splitting techniques. Because of the complexity of these computations and their large dynamic ranges, adaptive mesh refinement and related techniques find their way into almost all of these applications.

Just one example of this class of computation is supernova simulations, which are being performed, among other reasons, to gain confidence in using Type Ia supernovas as "standard candles" for cosmological distance measurements. Current 2D supernova simulations require roughly 1000 hours on the NERSC IBM Seaborg system for a single model, where the typical size of the atmosphere is 100 zones. We estimate that moving to full 3D simulations, using this same approach, would require cubing the number of zones. This will require a run of one million hours on a system with ten times the capability of Seaborg. Thus advanced adaptive mesh refinement techniques will be required to reduce the number of total zones to a tractable level. Assuming this can be done, we believe that with a 50 Tflop/s sustained system, we can begin to perform meaningful 3D computations.

Projections of future computational requirements for analysis of cosmic microwave background data are truly daunting. Analysis of WMAP data currently under way requires 3×10^{21} flop/s and 16 TB of memory. Planck data in 2007 will require 2×10^{24} flop/s and 1.6 PB of memory. CMBpol data in 2015 will require 1×10^{27} flop/s and 1 exabyte of memory.

2.7 BIOLOGY

Biology has heretofore not been a major consumer of NERSC resources, but several new projects are estimating greatly increased requirements over the next few years, especially the Genomes-to-Life (GTL) program. Among the challenges being addressed here is better understanding of how information in DNA creates, sustains, and reproduces living systems. Planned studies will focus on multimolecular assemblies ("molecular machines").

One 2004 INCITE project in biology aims to increase understanding of the complex processes which occur during photosynthesis. This is important to better understand carbon sequestration and energy transfer systems. A key code used here is a quantum Monte Carlo code, which is used to study systems involving hundreds of electrons.

A 2005 INCITE project has combined molecular dynamics and proteomics to create an extensive repository of protein fold structures. Hundreds of major computer runs are planned, using up to 1000 CPUs. Several times the current allocation will be required in the future.

2.8 CONCLUSION

We have sketched future requirements for several application areas. This is certainly not a complete list — we have not mentioned all application areas, and within each application area we have only listed one or two key applications. The common thread shared by these analyses is a greatly increased level of computational resources to meet future requirements. Indeed, a factor of 10 increase from current levels would merely serve the immediate backlog. Looking a few years into the future, we see requirements 100x and 1000x current capabilities. And some projects, such as computational biology, are just beginning to ramp up their consumption.

The large expected increase in experimental data, as typified by the cosmic microwave background project and genomics, will significantly change the landscape of high-end scientific computing. In the past, almost all data was generated by simulations. In the future, the balance likely will tip to experimental data. Developing the technology to store, search, analyze, and visualize this data will be a high priority.

Periodically, the NERSC User Group develops a "Greenbook" to express the computational requirements that the Office of Science will have over the next five to seven years. The requirements in the Greenbook are for the entire NERSC facility, not just a particular system. The general recommendations from the draft 2005 Greenbook are:

1. Expand the high-performance computing resources available at NERSC. The upcoming procurement must ensure a large increase in compute cycles available, as well as an appropriate balance of cache memory, processor memory, memory bandwidth, internode

communication speed, intranode communication speed, and other computational equipment required to support the wide range of large-scale applications involving production computing and development activities in the DOE Office of Science.

2. The computing hardware and queuing systems should be configured in such a way that minimize the time-to-completion of large jobs, as well as maximize the overall efficiency of the hardware. Here, "large" can refer to jobs requiring long running times, a large number of processors, exceptionally large memory, or any combination of these. We do not recommend policies that encourage individual jobs to use more processors than is justified by their memory requirements and scaling characteristics if it degrades overall facility efficiency and overall job time-to-completion.
3. NERSC should actively support the continued improvement of algorithms, software, and database technology for improved performance on parallel platforms. This implies that NERSC should continue to retain high quality consultants and performance support personnel to proactively assist the user community in obtaining maximum performance from the hardware.
4. Significantly strengthen the computational science "infrastructure" at NERSC that will enable the optimal use of the current and future NERSC supercomputers. Current simulations are generating data at tens of terabytes per simulation over the course of a few days. Simulations that will be performed in the not-too-distant future will generate hundreds of terabytes of data per simulation over the course of months. Moreover, these simulations will be performed, and the results analyzed, by geographically distributed teams. Consequently, local disk and archival storage, networking between NERSC's supercomputer and local storage and between NERSC and the WAN, and capabilities for local and remote data analysis and visualization must all be developed in order to enable the scientific "workflows" that will produce next-generation computational science.
5. Several of NERSC's current and potential future scientific applications are especially data or I/O intensive. These requirements should be carefully evaluated in order to support as wide a range of science as possible while also realizing significant benefits in both performance and cost in the computer configuration.

3. Technology Challenges 2006–2010

The peak performance of high-end computing platforms has increased by a factor of 1.8x annually over the past decade, with two trends contributing equally to this growth: (1) individual processor performance has grown by about 1.4x annually, due to both higher clock speeds and the increased use of on-chip parallelism; (2) the number of processors in high-end machines has increased by an average factor of 1.3x annually. Users of high-end systems have developed new algorithms and software techniques that allow their codes to take advantage of increased parallelism.

We expect these performance trends to continue in the next decade with two major differences. First, as opportunities for additional instruction-level parallelism wane, chip manufacturers will exploit smaller feature sizes by increasing the number of processors per chip. Indeed, some experts suggest that the number of processors per chip is likely to double every few years. Second, vendors of high performance computing systems will increasingly implement processor and network accelerators. These two factors will expose more parallelism to applications and potentially require significant code and algorithm changes to achieve high performance, placing an even higher burden on applications programmers.

At the same time, some of the most important large-scale DOE applications are moving to adaptive, unstructured, and tree-based algorithms. These methods tend to have smaller messages, more load imbalance, more asynchrony, and a higher ratio of communication to computation than their more regular predecessors. Ironically, these computationally efficient algorithms are some of the most challenging algorithms to map efficiently onto terascale systems. The algorithms in these applications exhibit large-scale parallelism, and often multiple levels of parallelism, but not necessarily to the degree or level required by the machines. Furthermore, while codes are generally written to be independent of the exact number of processors on which they run, the experience of NERSC users and others indicates that each order of magnitude increase in parallelism requires a significant rethinking of the algorithms and the parallelization strategy.

3.1 PROCESSOR TRENDS 2006–2010

The floating point performance of microprocessors as measured by the SPECfp benchmarks has improved by 59 percent per year over the 16-year period from 1988 to 2004. According to the National Research Council's *Getting Up to Speed* report,[2] an exponential trend in processor performance is expected to continue, but at a reduced rate. The increases in performance have come from technology improvements (Moore's Law), pipeline depth, and instruction-level parallelism.

Moore's law (the exponential decrease in feature size) is expected to continue through the next decade, but the increases in clock speed that we have come to associate with Moore's law may be smaller than in the past because of power density limitations. Increases in performance from increasing pipeline depth and instruction-level parallelism are not expected to continue. Instead, manufacturers will be placing several processor cores on a single chip. The end result is that chip performance may continue on an exponential curve, but with the numbers of processors per chip doubling every few years. This means that more of the on-chip parallelism will be the responsibility of the programmer, unless there are dramatic changes in architectural models or compilers that allow this level of parallelism to be hidden.

The gap between processor performance and memory system performance is already limiting the performance of many important scientific codes, and that gap will continue to widen over the next decade. The NRC report points out that in the period from 1982 to 2004, the bandwidth of commodity microprocessor memory systems increased by only 38 percent per year, and in the period since 1995, the rate has slowed to only 23 percent per year. The gap with memory latency is growing even more quickly, with latency improving by only 5.5 percent per year. NERSC needs to understand the impact of these gaps on the algorithms currently in use and how those algorithms will perform on architectures that evolve in the current models. Unfortunately, the more scalable algorithms such as multigrid tend to have lower ratio of floating point operations to memory operations; this low *computational intensity* leads to low single-processor performance.

[2] National Research Council Committee on the Future of Supercomputing, *Getting Up to Speed: The Future of Supercomputing,* S. L. Graham, M. Snir, and C. A. Patterson, eds. (Washington, DC: National Academies Press, 2004).

High end computing vendors are aware of the growing gap between peak performance and achievable performance for scientific applications. Many of these vendors are looking carefully at ways to accelerate performance while still leveraging all of the benefits of commodity processors. The DARPA HPCS program is one driver, but non-HPCS vendors are looking at accelerators as well.

A memory accelerator is one way to address memory bandwidth and latency issues. Memory accelerators such as hardware and software prefetch mechanisms will be used to hide memory latency. However, the hardware mechanisms are currently limited to unit stride memory access patterns and often to less than a dozen simultaneous streams. Software prefetch mechanisms may relax the unit stride limitation, but require that a compiler (or application programmer) can statically determine the memory access pattern. Meanwhile, codes based on 2D, 3D, and unstructured meshes tend to have strided and indexed memory access patterns, and codes that model complex physical systems may have dozens of arrays that are being read and written in a single loop nest.

Other types of processor accelerators are designed to increase instruction-level parallelism, which can help in hiding memory latency and also boost peak floating point speeds. All of the commodity microprocessors now have SIMD extensions, while the research community is looking at more general vector and streaming extensions to microprocessors. In the custom supercomputing market, vectors continue to have a presence in the Cray and NEC lines. FPGAs have also proven useful for some applications that need specialized instructions not available in commodity systems; while they support more general forms of parallelism, due to the need for configuration, they also tend to be used for data parallel algorithms. Finally, some of the highest performance processors are coming out of the high-volume game and graphics markets; these systems have been notoriously difficult to program, and results so far in the scientific arena are limited to a few small kernels, such as dense matrix multiplication.

It is likely that several high-end systems offered in the 2009 time frame will include processor accelerators of one type or another. NERSC needs to understand these processor accelerators and their effectiveness on a broad range of applications and algorithms in the NERSC workload. NERSC must work closely with vendors to ensure that the accelerators which vendors choose to implement will benefit the broadest subset of the NERSC workload. In general, these accelerators offer performance opportunities and software development challenges.

3.2 INTERCONNECT TRENDS 2006–2010

The gaps identified between processor and memory system performance have even wider counterparts in the interconnect arena. This has a direct impact on NERSC planning and procurements, as explained in the NRC report:

> The cost and power of providing bandwidth between chips, boards, and cabinets is decreasing more slowly than the cost and power of providing logic on chips, making the cost of systems bandwidth dominated by the cost of global bandwidth.[3]

[3] *Getting Up to Speed,* p. 117.

The problems of bisection bandwidth scaling, especially as the number of processors scale, has lead to an increasing interest in interconnect topologies with less than full bisection bandwidth, such as torus and related mesh topologies. For relatively small systems, such interconnects perform well, especially because point-to-point bandwidth is overprovisioned, resulting in reasonable bisection bandwidth. It has not yet been demonstrated that these topologies can be effective on NERSC's diverse workload at large system sizes (thousands of nodes). There are several reasons they might be less effective than a full bisection network: algorithms requiring high bisection, such as multidimensional FFTs, may perform poorly; application performance may vary depending on which processors it runs on; application performance may be significantly affected by other applications running on the system. To some extent these issues can be mitigated by aggressive process migration, but NERSC's experience on the Cray T3E and the lack of efficient migration software make it doubtful that this solution will be effective.

The gap between processor speed and network latency grows even more quickly than memory latency. Supercomputing network latencies have remained roughly constant in recent years, with user level latencies often dominated as much by software overhead and memory costs as by physical switch latencies. Machines with tighter integration have lower latencies, but often at a significant cost, since they may involve processor customization.

Various forms of network accelerators are being considered by researchers and HPC vendors to address the network performance gap. Support for direct access to remote memory has been available in custom supercomputing networks for many years, but at a high cost. The introduction of standard interconnects such as Infiniband has marked a trend in networking — these networks have a broader market than high performance computing, and still support direct access to remote memory. There is still a latency advantage to more highly customized networks that connect through a memory interface, rather than an I/O bus, but the performance difference is smaller.

The software overhead for message passing libraries is also significant, and vendors are responding by offloading some of the protocol processor onto the network interfaces. This may involve send/receive matching, remote atomic operations for synchronization, and support for accelerating collective communication. All of these factors lead to a complicated landscape of networking performance and cost characteristics, with application usage patterns being critical in procurements. In addition, while the standard interconnects offer hardware opportunities, they may not have software libraries that are as well tuned and supported as on customized system, thereby placing an increased demand on NERSC services to improve their performance.

3.3 System Software and Tools

The trend towards commodity hardware in the past decade has been matched with a trend towards open-source software. This has in turn shifted some of the software development burden from the vendors to the application programs and to user service organizations like NERSC.

For operating systems, the role of Linux has been increasing, and Linux may be ubiquitous on all high-end systems in the next five years, with micro-kernels used on the compute nodes of some systems. While that provides a certain level of uniformity across systems, which is useful in

system management and administration, Linux itself is also fragmenting, with RedHat, SuSE, Fedora, and others providing different versions for which compatibility is a significant issue. HPC vendors are now providing horizontal solutions, with hardware and software components obtained from many different sources, rather than vertical solutions, in which a single vendor develops the entire hardware and software stack. Since there is no longer a single vendor for the system an all related software, facilities like NERSC must integrate software from diverse sources that has not been thoroughly tested.

In the last ten years the parallel computing community has standardized on MPI. Serial code is almost always written in Fortran, C, C++, or a mixture of these, sometimes with OpenMP extensions for SMP nodes. Scripting languages such as Python have shown some increase in popularity for gluing together application components. There is considerable concern at NERSC and in the parallel computing community as a whole as to whether the MPI model will be able to sustain scientific computing through 2010. MPI adds significant overhead, particularly in latency, and effectively constrains applications to a bulk synchronous programming model. C and Fortran do not have good mechanisms for expressing parallelism. The result is that MPI/C/Fortran applications may have difficulty taking advantage of the processor and network accelerators described above. Moreover, as scientific applications become more complex, the MPI model becomes increasingly difficult to work with. Phillip Colella explained this eloquently in the *Getting Up to Speed* report:

> Success in computational fluid dynamics has been the result of a combination of mathematical algorithm design, physical reasoning, and numerical experimentation. The continued success of this methodology is at risk in the present supercomputing environment, due to the vastly increased complexity of the undertaking. The number of lines of code required to implement the modern CFD methods such as those described above is far greater than that required to implement typical CFD software used twenty years ago. This is a consequence of the increased complexity of both the models, the algorithms, and the high-performance computers. While the advent of languages such as C++ and Java with more powerful abstraction mechanisms has permitted us to manage software complexity somewhat more easily, it has not provided a complete solution. Low-level programming constructs such as MPI for parallel communication and callbacks to Fortran kernels to obtain serial performance lead to code that is difficult to understand and modify. The net result is the stifling of innovation. The development of state-of-the-art high-performance CFD codes can be done only by large groups. Even in that case, the development cycle of design-implement-test is much more unwieldy and can be performed less often. This leads to a conservatism on the part of developers of CFD simulation codes: they will make do with less-than-optimal methods, simply because the cost of trying out improved algorithms is too high. In order to change this state of affairs, a combination of technical innovations and institutional changes are needed.[4]

Interest in languages with features to support high performance computing has been increasing over the last few years. These include UPC and Co-Array Fortran, as well as three new languages being developed as part of the DARPA High Productivity Computing Systems (HPCS) project. It is possible that one or more of these languages could reach critical mass sometime in the next ten years. NERSC should play an active role in understanding which languages and features are most important to the NERSC user community, providing feedback to the language community and, should any of the approaches prove promising, providing support to users in adapting their codes to the new languages. As *Getting Up to Speed* explains, the

[4] *Getting Up to Speed,* p. 154.

success of any major shift of this kind requires a complex "ecosystem," with architectures, software tools, and users all lined up behind the solution, and with facilities like NERSC central to such an ecosystem.

4. Science-Driven Systems

4.1 NERSC's Response to Technology Challenges

The years 2000–2005 were characterized by increasing standardization within the high performance computing community. Most high performance computing (HPC) codes, including almost all codes used at NERSC, are based on MPI with standard C, C++, or Fortran (sometimes with OpenMP), and large investments were made in developing and improving codes using this model. This software standardization accompanied a hardware standardization — most HPC systems are clusters based on superscalar processors with high performance interconnects well suited for MPI. Systems that can support different models are often used with the standard MPI code.

While it is possible that this model will continue, the technology changes described earlier will present several opportunities and challenges:

- New systems will have much more parallelism than today's systems because of the introduction of multicore CPUs. Enormous resources (e.g., SciDAC) have been expended to scale up existing applications, and thus there is less "low hanging fruit" than before to achieve increased parallelism.
- New system architectures will include new processor architectures and heterogeneous processor architectures. These new processors are a different approach to dealing with the multicore "problem" above.
- A relative increase in latency to local and remote memory will be coupled with the availability of new latency hiding techniques through language and hardware improvements.
- Mesh-based networks may become more readily available at a lower cost.

Depending on the types of changes, it may take several years to adapt a scientific code to a new architecture or programming paradigm. Whether or not it will make sense to adapt NERSC codes to meet these new challenges depends on both the commercial availability of the new hardware and software and on the potential performance and productivity improvements. NERSC will address these questions proactively, rather than reactively, by continuing the Science-Driven Architecture initiative, in which NERSC will work with computer vendors to develop new systems that address the needs of scientific applications in general and the NERSC workload in particular.

There are four major areas of system design and implementation at NERSC: the computational systems; the storage systems (online and nearline); the networking, analytics, and visualization systems; and servers for supporting functions. The balance of the entire Center is determined by the requirements that evolve from the increased computational capability, plus independent requirements for other resources. New systems must be designed to support not just current

work, but future workloads as well. The following paragraphs provide detailed descriptions of NERSC's strategy for its computational systems, storage systems, and network.

4.2 COMPUTATIONAL SYSTEMS

The NERSC computational system strategy for increasing researchers' productivity is to provide high-end systems — NERSC focuses on the balanced and timely introduction of the best new technologies for complete computational and storage systems. The NERSC plan is designed to maintain a balanced combination of systems and services to address the widest breadth of DOE capability science. The basic strategy NERSC has followed will continue — having two major computation systems in place at a time — with modest-sized systems arriving in between the major systems as funding and technology allow. Every three to four years the oldest generation system is replaced with the latest generation system. A three- to four-year interval is based on the length of time it takes to introduce large systems, the time it takes for NERSC clients to become productive on new systems, the rate of change of computational technology, and the types of funding and financial arrangements NERSC uses.

Normally, NERSC has two generations of major computational subsystems in service, so that each subsystem will have a lifetime of five to six years. This overlap provides time for NERSC clients to move from one generation to the next, and provides NERSC with the ability to fully test, integrate, and evolve the latest generation while maintaining service on the earlier generation.

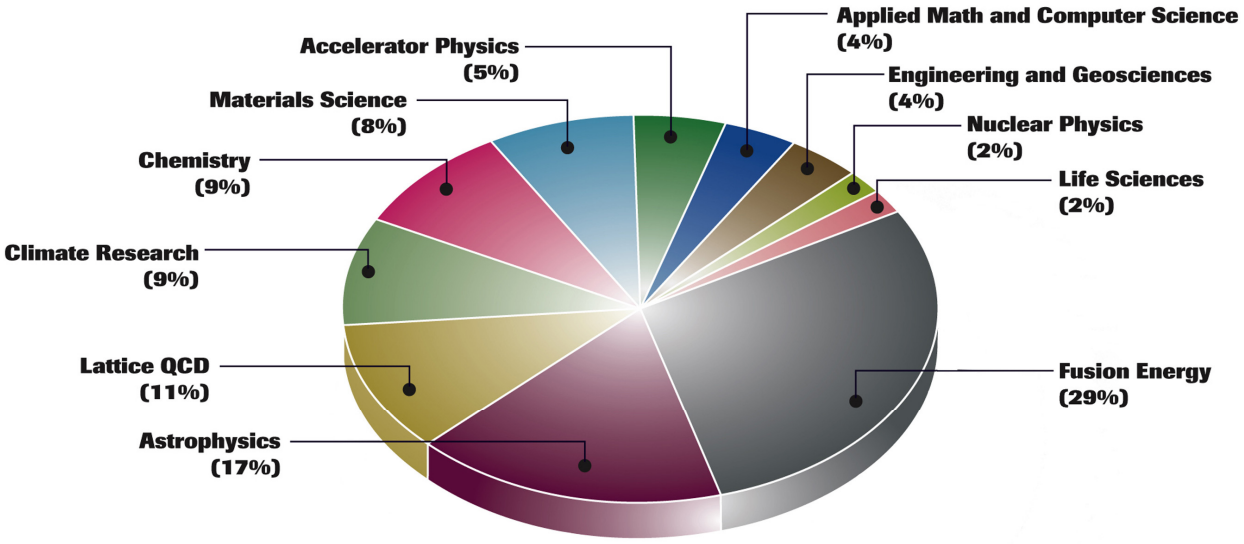


Figure 6. NERSC usage by scientific discipline in 2004 shows that the NERSC workload includes almost every type of science.

NERSC must support a diverse workload, as shown in Figure 6. NERSC's highest priority is to support *capability computing*, which we define here as the need to use more than one-tenth of an entire computing resource over an extended time period. This category includes INCITE and SciDAC projects. NERSC also supports *large-scale computing*, which is defined as the use of significant computational resources over an extended time period. Finally, NERSC supports a

small amount of *related capacity computing*, which is comparable to running on a desktop system for a week. The National Research Council report on supercomputing, *Getting Up to Speed: The Future of Supercomputing,* defines capability and capacity computing as follows:

> The goal [of capability systems] is to solve a large problem or to solve a single problem in a shorter period of time. Capability computing enables the solution of problems that cannot otherwise be solved in a reasonable period of time (for example, by moving from a two dimensional to a three-dimensional simulation, using finer grids, or using more realistic models). Capability computing also enables the solution of problems with real-time constraints (e.g., intelligence processing and analysis). The main figure of merit is time of solution. Smaller and cheaper systems are used for capacity computing, where smaller problems are solved. Capacity computing can be used to enable parametric studies or to explore design alternatives; it is often needed to prepare for more expensive runs on capability systems. Capacity systems will often run several jobs simultaneously. The main figure of merit is sustained performance per unit cost. There is often a trade-off between the two figures of merit, as further reduction in time to solution is achieved at the expense of increased cost per solution; different platforms exhibit different trade-offs. Capability systems are designed to offer the best possible capability, even at the expense of increased cost per sustained performance, while capacity systems are designed to offer a less aggressive reduction in time to solution but at a lower cost per sustained performance.[5]

NERSC uses these definitions in this plan.

4.3 COMPUTATIONAL TECHNOLOGY CHOICES

There are three basic types of technology in this time frame that need to be considered, as defined by the National Research Council study:

- *Commodity supercomputers* built using off-the-shelf processors developed for workstations and commercial applications and connected by off-the-shelf networks such as Ethernet that connect to the nodes via I/O buses. Examples are NERSC's PDSF system, the Virginia Tech Big Mac cluster, the UCB Millennium cluster, and a host of integrated solutions.
- *Custom supercomputers* using processors and interconnects that are specialized for scientific computing. The systems provide specialized and high bandwidth interconnects and processor-memory interfaces. Examples are the Cray X1 and the NEC SX-8.
- *Hybrid supercomputers* combine commodity processors with custom high-speed interconnects. Examples include the ASCI Red Storm, Cray T3E, SGI Altix, and IBM SP. Hybrid systems may also have semi-custom node configurations (e.g., the Power 5 "Blue Planet" node), special value-added accelerators (e.g., ViVA), and special packaging.

The NERSC workload it too diverse to focus exclusively on either the commodity supercomputer or the custom supercomputer. Subsets of the NERSC workload can operate effectively on either architecture, but neither type of system is sufficient. Hence NERSC has three strategy choices.

The first strategy is to subdivide its computational investment into multiple smaller systems that focus on the subset of the workload that operates best on that architecture. This approach has advantages such as having the codes run on the most efficient architecture. It has several

[5] *Getting Up to Speed,* p. 24.

disadvantages, one of which is that subdividing the resources prevents being able to do capability and large-scale computing. Another disadvantage is the loss of long-term flexibility. Over the past four years, the NERSC workload has changed both in the amount of allocation going to different disciplines and in the use of different algorithms within a discipline. Custom and completely commodity systems may not be able to adapt to these changes in a timely manner.

The second system strategy is a variant of the multiple system strategy called "evergreen" since hardware is augmented regularly, with the new hardware replacing the oldest. This has the advantage of being able to include the latest and/or lowest-cost technology into a system, but has a significant limitation in that the system becomes heterogeneous. An example of this is how the NERSC maintains the PDSF system. While some of the workload can effectively take advantage of this approach, almost all of the capability workload is bound by Amdahl's law, which requires all the system components to be homogeneous. Until the capability applications can adapt their load balancing to accommodate different-speed CPUs, the evergreen strategy is not feasible for the major computational systems.

The third strategy, the one NERSC selects, is to focus on hybrid supercomputers; to use NERSC's *science-driven computer architecture* activities to improve the hybrid computers to run the entire NERSC workload well; and where possible, to add additional functionality such as ViVA that will accelerate some or all of the NERSC workload. The hybrid approach (Figure 7) has several advantages:

- NERSC has demonstrated the ability to operate such systems in both the capability and capacity mode. Thus, NERSC can rapidly respond to the changing needs of the DOE Office of Science and the national research base.
- The systems cover a wide range of discipline areas and methods and thus provide excellent flexibility for changing allocation priorities and algorithmic methods.
- The systems cover a wide range of discipline areas and methods and thus provide excellent flexibility for changing allocation priorities and algorithmic methods.
- The system will be easier for users to use and system managers to support because dramatic programming changes do not occur between systems as they would with the first two strategies.

The total annual investment in the computational systems will be approximately one-third of the total NERSC annual funding. As in the past, lease-to-own payments for a system will be spread over three years, and it is possible that technology availability will dictate a phased introduction over one year to 18 months.

NERSC will use the "best value" process for procuring its major systems. This 22-step process allows considerable flexibility for NERSC and also provides an opportunity for significant innovation by suppliers. The principal task of the acquisition team is to decide the best alternative among the available choices. One key metric we use is what we call the Sustained System Performance (SSP) metric, which is based on a benchmark performance integrated over three years. We will use the Effective System Performance (ESP) test to assess system-level efficiency, namely the ability of the large-scale system to deliver a large fraction of its potential resources to the users. In addition, NERSC will use a sets of benchmark kernels and full

applications to assess system. NERSC will also use advanced modeling methods for its applications to project the performance of systems as well as to guide the Science-Driven System Architecture efforts.

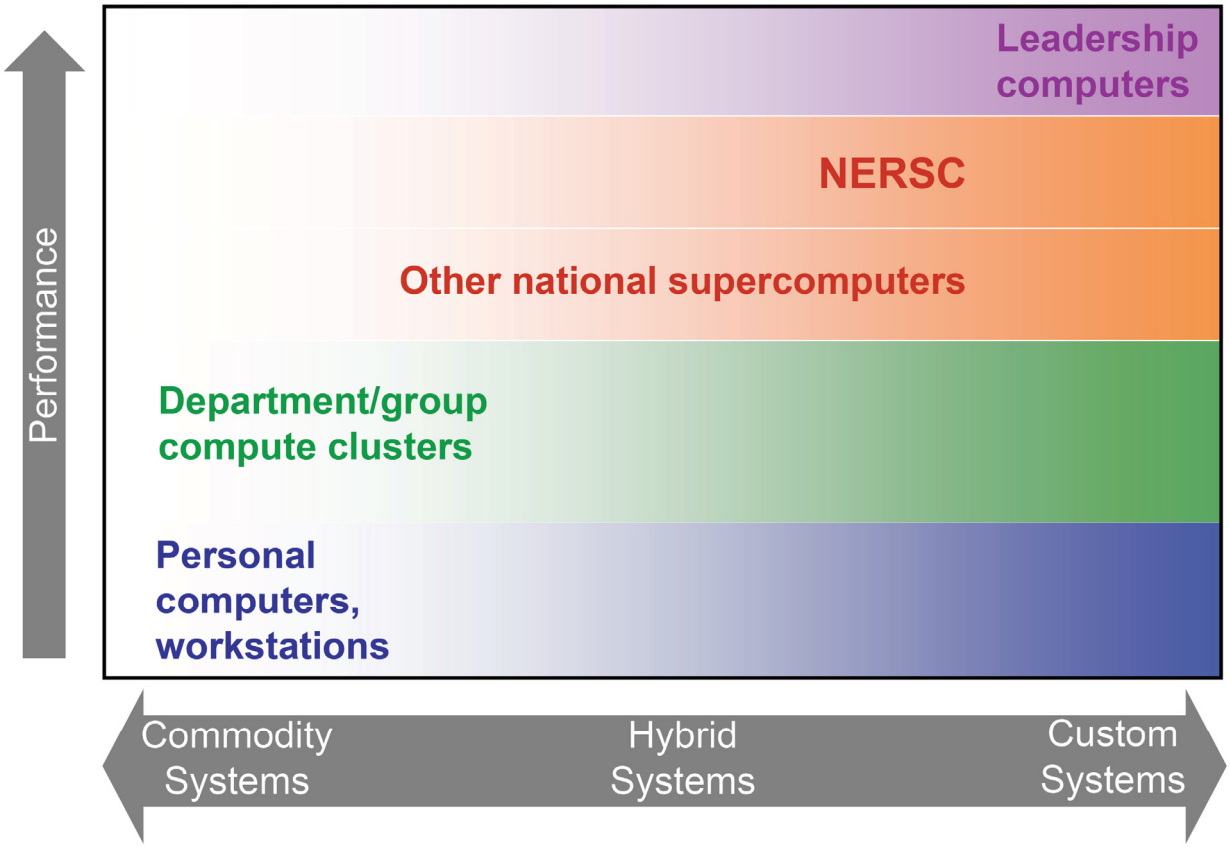


Figure 7. From the spectrum of scientific computing systems, NERSC's strategy is to acquire high-end hybrid systems.

4.4 Computational System Sizing

This section intentionally left blank—proprietary information.

4.5 Facility-Wide File System

In FY2005, NERSC began deployment of a Facility-Wide File System (FWFS). FWFS will provide consolidated storage for online user data at NERSC, replacing traditional system-local parallel file systems for home directories, scratch storage, and project storage. While FWFS is external to all computational systems, it will be mounted natively and at high performance. FWFS will grow and evolve over time, serving several generations of computational systems.

The benefits of FWFS to scientific productivity are manifold. By providing a single unified namespace, FWFS will make it easier for users to manage their data across multiple systems. Users will no longer need to keep track of multiple copies of programs and data; they will no longer need to copy data between NERSC systems for pre- and post-processing. Storage utilization will become more efficient through decreased fragmentation. Computational resource utilization will become more efficient as users can more easily run jobs on an appropriate resource. Storage allocations (quotas) will become larger, because they will no longer be fragmented among several systems. FWFS will also provide improved methods of backing up user data that has the potential to mitigate disturbance to users when block backups are run.

FWFS is a part of NERSC's strategic investments in analytics, described in Section 6 below.

Anticipated developments in file system technology will provide further benefits. It is expected that FWFS will provide Hierarchical Storage Management (HSM) functionality, in which data can be automatically migrated to and retrieved from tertiary storage. This will further improve scientific productivity by enabling the user perception of extremely large online disk (very large quotas) and making it unnecessary for users to manually transfer data to and from HPSS. Other possible technology enhancements include enhanced security, wide area access, and tighter integration with grid technologies.

The initial FWFS deployment will be a modest step, supplementing local parallel filesystems rather than replacing them, but giving users the opportunity to adapt their workflow to take advantage of the functionality provided by a centralized file system. NERSC has selected the GPFS filesystem for this deployment, and is working closely with IBM to ensure that GPFS will be available on future NERSC systems, and that GPFS follows an aggressive roadmap to further improve functionality and performance. As new systems are brought in to NERSC, these systems will be more tightly integrated with FWFS. By 2010, it is expected that FWFS will be available on all potential NERSC production systems.

4.5 MASS DATA STORAGE

Science-driven data storage involves archiving and enhancing access to the large amounts of new scientific data as well as large historic repositories of scientific data. NERSC will scale the storage system to handle the amount of data expected as the Center evolves, which includes frequent upgrades for storage technologies while keeping the footprint for storage systems as small as possible, configuring those systems for efficient data access, and moving the archive to new network and processor technologies as the Center infrastructure changes.

The NERSC archive currently holds 1.5 petabytes (PB) of data, and archival capacity will be increased to 16 PB this year. The archive is expected to grow to nearly 40 PB over the next five years (Figure 8). Projections of data growth are used to plan and scale the system, including adopting new storage technologies and planning data placement and the storage system footprint. Five-year projections are used for planning gradual improvements and technology changes — for example, a new technology density is required by 2009 to grow archival capacity — while actual data growth is used to adjust the plan yearly. We have developed a path forward that optimizes the use of our media and new technology deployment to keep the archive dense and affordable.

NERSC's plan addresses the cost of media that is consumed and keeps the cost of media (purchased tapes) essentially flat rather then reflecting the exponential growth of data.

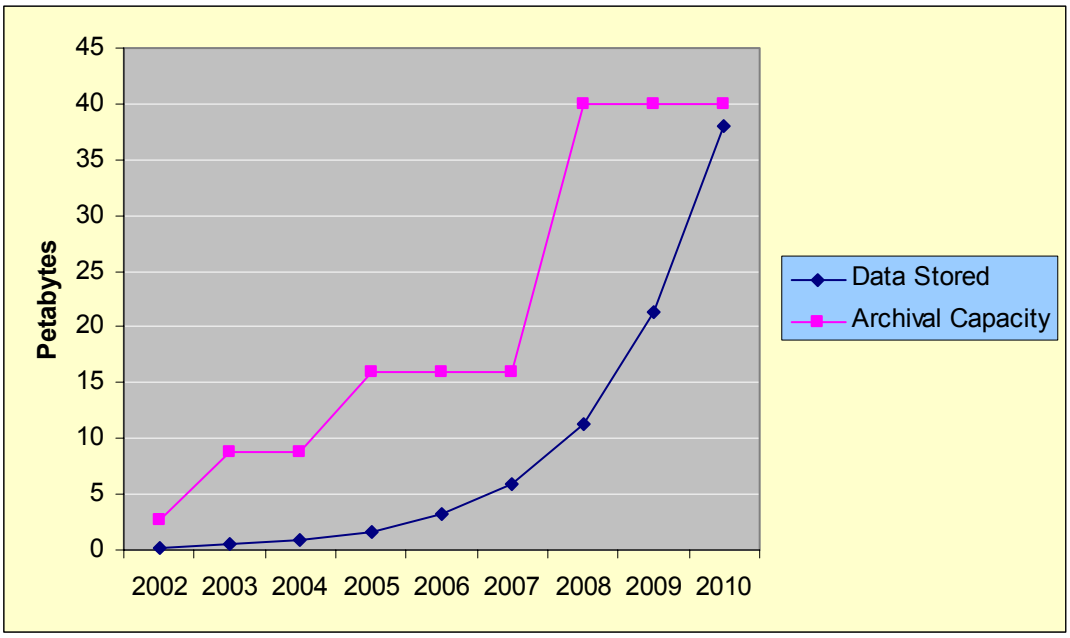


Figure 8. Projected growth of stored data and archival capacity.

These projections are also used to plan optimal access to data. Table 2 shows projections of minimal data access requirements based on projected data growth. Storage system configurations are changed to map to the expected transfer rate requirements. When the storage media will not sustain the transfer requirements, we configure parallel data transfers (stripes) to meet the bandwidth requirements. When data bandwidth requirements exceed the network bandwidth to storage, data is striped to multiple storage nodes using multiple network interfaces.

Table 2
Projected Data Growth and Bandwidth Requirements, 2005–2010

Year	Total Archived Data	Data Transfers per Day	Transfer Rate
2005	1.5 PB	6 TB	60 MB/s
2006	2.9 PB	10 TB	120 MB/s
2007	5.0 PB	20 TB	231 MB/s
2008	11.0 PB	33 TB	392 MB/s
2009	21.0 PB	64 TB	749 MB/s
2010	38.0 PB	117 TB	1356 MB/s

NERSC storage systems are currently capable of single transfers to disk of 200 megabytes/sec (MB/s), with a planned upgrade to 400 MB/s this year. On the surface, this would seem to be plenty of bandwidth for 6 TB peak days; however, Figure 9 shows that actual activity to storage currently has peaks of 500 MB/s. This year's single-stream capability of 400 MB/s will generate

1 Gb/s peaks, taking advantage of NERSC's upgrade to 10 Gigabit Ethernet this year and also allowing for peak days of 33TBs.

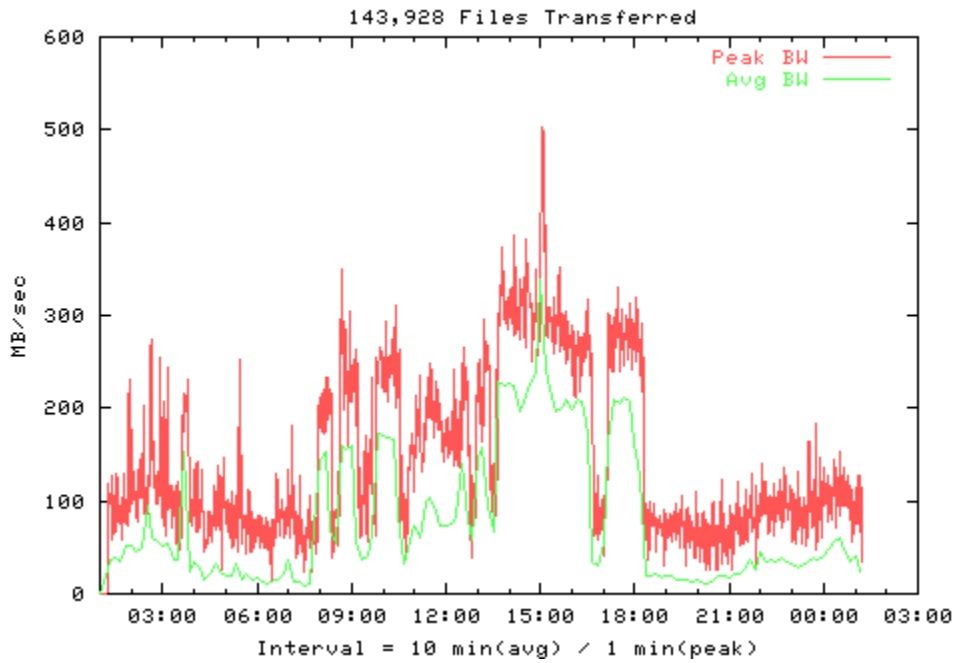


Figure 9. NERSC aggregate bandwidth for November 5, 2004.

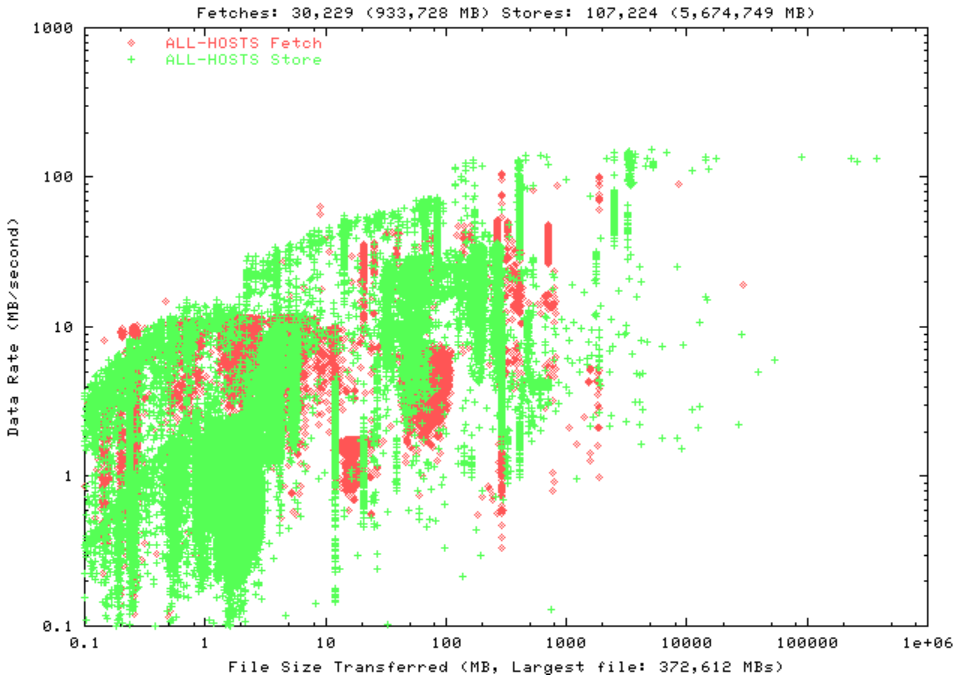


Figure 10. Size of data files transferred on November 5, 2004. The red points are the files retrieved from the HPSS archive, and the green points are the files transferred into HPSS for storage.

The size of individual files has grown tremendously over the past few years. Figure 10 is a dot plot of individual transfers to storage for one day. File sizes over 1 GB are common (1000 on

this plot). The largest file transferred this day was 372 GB. Configuring the storage system for good data access requires considering individual file transfer bandwidth. Before HPSS was deployed, NERSC's storage systems were constrained by the slowest resource: media, processors, or network. With HPSS we have been able to configure around resource bottlenecks. For instance, current tape technology streams data at only 30 MB/s. At this rate, a transfer of a 372 GB file would take 3 to 4 hours. HPSS is highly configurable, and striping the data across three tape drives can stream the data at 90 MB/s, shortening read access time to this 372 GB file on tape to 1 hour. This year NERSC will invest in new technology tape drives with transfer rates of 120 MB/s. Using the current stripe, access time will be 15 minutes to this 372 GB file, and transfer rates to striped tape will be 360 MB/s. Data stripes have also been used to increase the network bandwidth to storage: using a four-way disk stripe configuration and multiple network interfaces, the network speed to data was increased from 80 MB/s to 200 MB/s. With the upgrade to 10GigE, each individual stripe of data will be capable of full disk bandwidth, raising transfer capability to 800 MB/s.

NERSC is one of the original HPSS development partners, along with LLNL, LANL, SNL, ORNL, and IBM. NERSC is the only site where this effort is funded by Office of Science funds, yet HPSS is critical to a number of SC projects that are independent of NERSC. Further, HPSS is the only archive storage system that can meet the scalability, performance, and feature requirements for capability computing. NERSC will continue to participate in the HPSS collaboration, which brings a number of benefits, including early access to new features and new technology, ability to prioritize requirements and strategy, and collaboration with NNSA activities.

In addition to handling the amount of data growth, keeping the footprint reasonable, and optimizing access to data storage, storage strategies center around data safety, data availability, and data preservation over time. These areas are critical requirements for large archives.

Data safety is implemented on many levels: catch it, save it, keep it. Our first level of hierarchy is high performance HPSS cache RAID5 and RAID3 disk. Data is written to tape immediately after it arrives on HPSS cache disk. Our tape environment is fully automated, and storage tape silos are strong, resistant to damage, and equipped with fire suppression. The storage system metadata is backed up offsite for disaster recovery every three months.

To meet availability requirements, the storage system is configured using multiple redundant components to minimize the effect of failures of any one component on the availability of the whole storage system. An outage of any device or system does not cause a storage system outage.

The NERSC storage systems hold 30 years of archival data. All of this data has been moved to new storage systems and new hardware, and is currently available nearline. NERSC gradually introduces new technology alongside existing technology, allowing for its reliable introduction. This methodology and configuration has been used not only to move all of the data in storage to new technology devices but also to completely replace our underlying network, devices, and processors multiple times. Being able to move both the storage system and the data to new technology results in the active preservation of data repositories over long periods of time.

Over the past five years, HPSS has fully scaled to the NERSC Center's need for data storage. Hardware upgrades for FY 2005 will replace our bonded internal and external Gigabit Ethernet networks with 10 Gigabit Ethernet internal and external networks, and will add dense tape drives, increasing the capacity of the archive to 16 PB and the bandwidth to tape to 120MB/s. Software upgrades this year will replace the HPSS server infrastructure and replace the metadata database, significantly improving access to HPSS metadata. Software upgrades in FY 2006 will replace the HPSS security infrastructure and add features such as native Grid capability and possibly a Facility-Wide File System/Hierarchical Storage Management system.

4.6 NETWORKING

As a national facility with users who routinely transfer massive datasets of 100 GB or more, NERSC needs fast and scalable networks as much as fast and scalable computers. The local area network (LAN) within NERSC and the wide area network (WAN) external to NERSC must allow NERSC users to easily and efficiently access NERSC resources and transfer data to, from, and within NERSC.

NERSC currently averages 2 TB of WAN traffic per day (~185 megabits/sec [Mb/s]), while heavy throughput days approach 4 TB/day (375 Mb/s), as shown in Table 3. Peak periods through the day are roughly twice the daily average (400–750 Mb/s). To sustain high transfer rates for large datasets, it is important to avoid triggering the automatic congestion control mechanisms in TCP (the Transmission Control Protocol), which may stall the transfer and require hours to recover to the full transfer rate. NERSC has observed that in most cases, bandwidth with 2x "headroom" above the peak rate is sufficient to avoid transfer stalls. Therefore, NERSC provides bandwidth 4x above the daily average — 2x for peak and 2x for headroom.

Table 3
Wide-Area Network Traffic

	Data Moved	Average Rate Required	Peak Rate Required	Headroom to Support Large Transfers
Current average day	2.0 TB	185 Mb/s	400 Mb/s	800 Mb/s
Current heavy day	4.0 TB	375 Mb/s	750 Mb/s	1.5 Gb/s
Next upgrade point	6.5 TB	600 Mb/s	1.2 Gb/s	2.4 Gb/s

Effective bandwidth on NERSC's existing OC-48 (2.4 Gb/s) ESnet link will begin to suffer when 4x the average daily transfer rate approaches the link speed. That threshold would be an average rate of 600 Mb/s, or approximately 6.5 TB per day. Historically, data stored in NERSC's HPSS increases at the rate of 1.7x per year, as shown by the red line in Figure 11; the blue line shows the NERSC–ESnet link speed over the same time interval. Assuming the 1.7x rate of increase for archive data continues, NERSC is likely to reach 6.5 TB/day around October 2005. To minimize the chance of impacting scientific data transfers, the NERSC–ESnet link will be upgraded to OC-192 (10 Gb/s Ethernet) before then. If the trend holds, the NERSC–ESnet link will need to be upgraded to 40 Gb/s by August 2008 and 80–100 Gb/s by 2010.

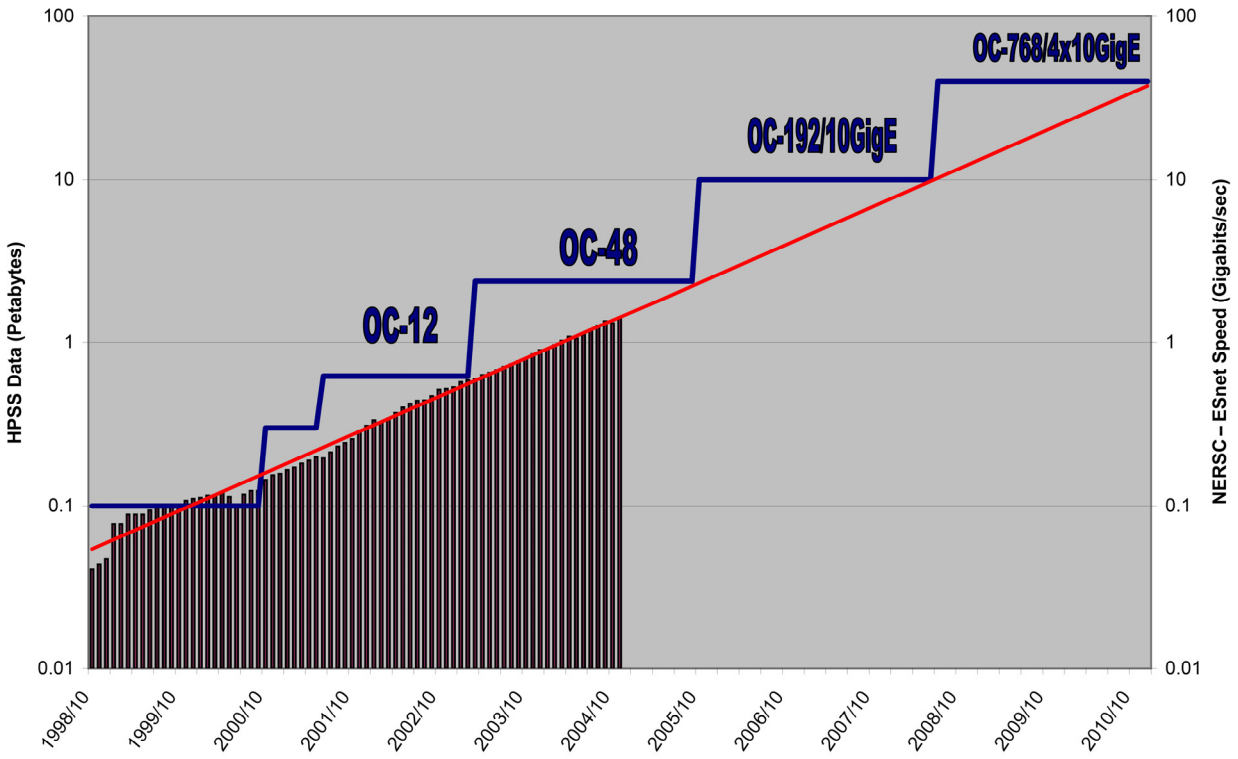


Figure 11. NERSC data storage and WAN network speed, 1998–2010.

To enhance both network reliability and bandwidth to NERSC and other DOE research sites in the San Francisco Bay Area, ESnet will complete the Bay Area Metropolitan Area Network (BA MAN) in the summer of 2005. The BA MAN will be constructed from two rings of fiber around San Francisco Bay and will provide multiple 10 Gb/s paths between NERSC, other Bay Area scientific facilities, and the WAN. NERSC plans to connect to the BA MAN during the summer of 2005 with one 10 Gb/s link for production traffic and a second 10 Gb/s link for special projects, dedicated high-bandwidth services, and testing.

As the demand for network resources grows, the NERSC network will continue to be expanded to match the scale of new systems and the performance improvements of the production network backbones. Additional 10 Gb/s production links can be easily and quickly added to the BA MAN, and we anticipate converting to 40 Gb/s links when the technology becomes cost-effective.

The NERSC LAN is designed as two separate networks: an internal network that provides optimized high-speed links between NERSC computing, visualization, and storage resources, and an external network that is optimized for transfers to NERSC user sites. In 2004 NERSC started upgrading the internal network to 10 Gb/s, and this expansion will continue in 2005 and 2006. Also in 2005, NERSC will procure new router and switching equipment to upgrade the NERSC external network to 10 Gb/s to connect to the BA MAN.

At the same time, NERSC is working closely with ESnet to create network peering arrangements that will maximize the effective remote access to NERSC users regardless of their institutional

affiliation and facility location. In particular, ESnet and Abilene are implementing high-speed peering between their networks at each of their collocated hubs at Sunnyvale, Chicago, New York, and Atlanta to create a common network backbone that provides very high-speed connectivity between NERSC and other labs and universities, comparable to what either backbone alone can provide among their primary sites.

In addition to its close collaboration with ESnet for production network infrastructure, NERSC will continue to consider direct connections to major experimental and dark fiber networks, such as the TeraGrid, DOE Ultranet, and National Lambda Rail, in order to support the computer science and network research community that combines sensors, archival data, and supercomputers to accomplish large multidisciplinary scientific projects. Once these advanced networks are in place, NERSC can assist "early adopter" scientific projects in using the networks. Connection to these experimental networks makes sense if it is done several years before NERSC needs to deploy the bandwidth in production. Hence we would look to participate if the aggregate network performance were 4x what the production network needs at any given time. Connections of this type would require additional funding for startup and ongoing costs.

As one example of NERSC's close collaboration with ESnet, we are working together on a project that will deploy a Quality of Service (QoS) capability that will permit high-priority network traffic to receive dedicated bandwidth. This capability will initially operate between ESnet sites and is expected to be expanded within two years to allow dynamic provisioning of circuits across both Abilene and ESnet. These "bandwidth corridors" could support real-time processing of experimental data at NERSC between experimental runs at remote facilities.

NERSC employs its network expertise to troubleshoot and optimize data transfers between remote user sites and NERSC resources. NERSC is one of the few sites that will troubleshoot problems end to end and actively engage with ESnet, Abilene, and other Internet service providers to determine exactly why a NERSC user is unable to utilize a significant fraction of the network's bandwidth.

4.7 VISUALIZATION AND ANALYSIS SYSTEMS

The primary aim of NERSC's visualization and analysis system strategy is to provide computational facilities especially well suited to the needs of data-intensive analysis applications, with interactive visualization and analysis — analytics — being the programmatic focal point. The architectural balance for data-intensive applications places more emphasis on I/O and large memory than on raw computational power. Over the years and into the foreseeable future, the most suitable systems for data-intensive computing are scalable shared-memory systems. In the late 1990s and early 2000s, the SGI Onyx platform was the best choice for providing raw I/O bandwidth, a scalable shared-memory architecture, support for diverse parallel programming models, and cache-coherent non-uniform memory access (ccNUMA). To complement the data-intensive architecture and user activities, the center has provided high-speed secondary scratch storage that is directly attached to the data-intensive platform.

The data-intensive architecture bias is required to meet the needs of the visualization and analysis workload. Especially in the case of interactive work, the load on the machine goes from

near idle to 100% capacity across all processors and I/O channels in response to the "bursty" nature of interactive visualization. Our visualization software offerings are selected in part to take advantage of such an architecture to provide the greatest possible capabilities to the NERSC user community. The shared-memory, ccNUMA architecture is especially well suited to serial analysis jobs that require access to vast amounts of memory: on an SMP ccNUMA architecture, a single process has access to the entire system memory. No other machine at NERSC offers such capabilities. Data-intensive applications, and visualization in particular, are characterized by the need to quickly load vast amounts of data from secondary storage, perform processing, and generate imagery. Such a pattern results in an exorbitant I/O load, and SGI systems are currently best suited to providing such balance.

Over the past few years, NERSC has made significant progress in better integrating its data-intensive analysis platform into the fabric of the Center. The present platform is included on the same internal 10 Gigabit network as the HPSS storage system. Whenever possible, licenses for commercial applications — especially visualization applications — are managed from a central location, which results in streamlined maintenance operations and better service for the user community.

Looking into the future, we have several expectations for the visualization and analysis systems at NERSC. First, we expect continuation of activities that aim to better integrate the visualization and analysis platform with the rest of the center. Such activities produce positive benefits for the users, and have a positive impact on the center itself in terms of better communication and efficient interaction between the groups.

Several activities are planned for the near future that continue these trends. The new Facility-Wide File System (FWFS) will be deployed to several platforms, including Seaborg and the new visualization and analysis system, DaVinci. A longstanding request from the user community is the presence of a high performance shared file system on both the computational and analysis platforms. Migration from /etc/passwd authentication on the visualization platform to a model that uses LDAP will enable users to use their NIM username and password to access the visualization and analysis platform, thereby making the system more easily accessible.

We will continue to expand the capabilities of the visualization and analysis platforms (as budgets permit) to meet the rising tide of computational data being processed at NERSC. In the upcoming year, we anticipate quadrupling the size of the visualization and analysis platform from an 8 CPU, 48 GB system to one with 32 CPUs and 192 GB. The increased platform capacity helps address an impedance mismatch between the large data sizes currently being produced and what is required to perform effective visual analysis. Further capability expansions are planned in 2007 and 2009.

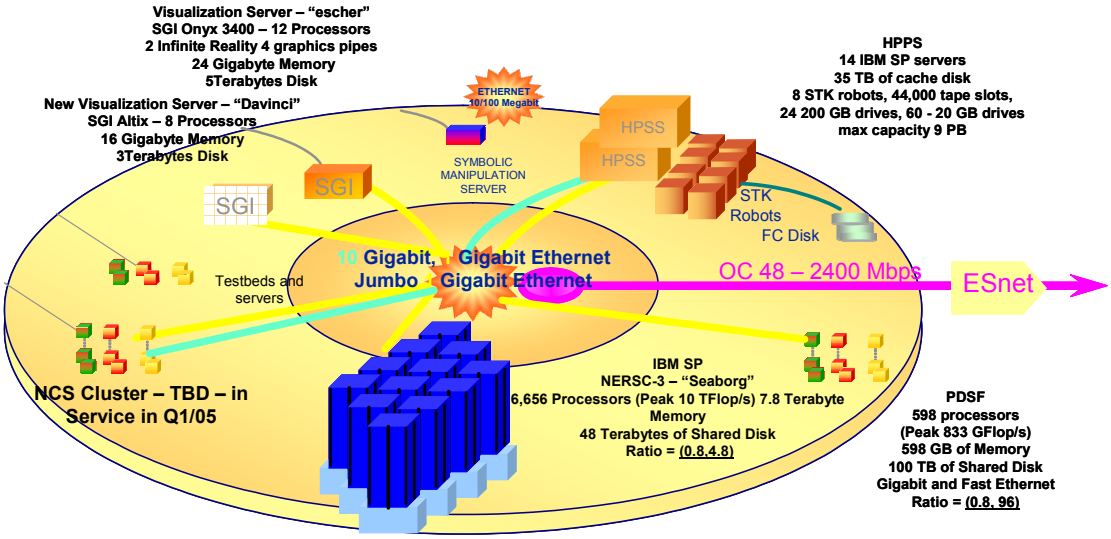
4.8 SCIENCE-DRIVEN SYSTEMS SUMMARY

The summary of the NERSC plan is:

2005

(Shown in Figure 12)

- NCS enters full service.
  - Focus is on modestly parallel and capacity computing.
  - >15–20% of Seaborg
- WAN upgrade to 10 Gb/s
- Upgrade HPSS to 16 PB. Storage upgrade to support 10 GB/s for higher density and increased bandwidth.
- Quadruple the size of the visualization/post-processing server.



Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Figure 12. NERSC's science-driven systems in 2005.

2006

- NERSC-5: initial delivery with possibly a phasing of delivery.
  - 3 to 4 times Seaborg in delivered performance
  - Used for entire workload and has to be balanced
- NCSb enters full service.
  - Focus is on modestly parallel and capacity computing
  - >30–40% of Seaborg
- Replace the security infrastructure for HPSS and add native Grid capability to HPSS
- Storage and Facility-Wide File System upgrade.

2007

- NERSC-5 enters full service.

- Storage and Facility-Wide File System upgrade.
- Double the size of the visualization/post processing server.

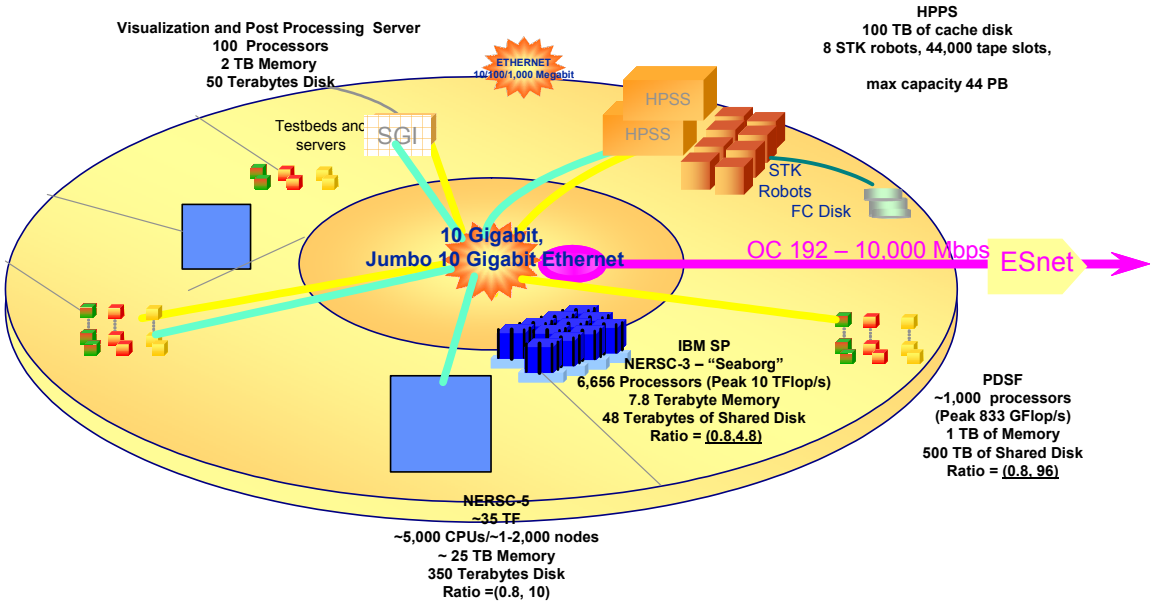
2008

- NCSc replaces NERSC-3/3E and NCS.
  - Uses maintenance money for N3E to replace NERSC-3 and NCS with a new capacity system that will be approximately the computational power of the combined systems.
- Increase archive capacity to 40 PB.
- Facility-Wide File System upgrade.
- Upgrade WAN to 40 Gb/s.

2009

(Shown in Figure 13)

- NERSC-6: initial delivery.
  - 3 to 4 times NERSC-5 in delivered performance
  - Used for entire workload and has to be balanced
- NCSc enters full service.
- Storage and Facility-Wide File System upgrade.
- Double the size of the visualization/post processing server.



Ratio = (RAM Bytes per Flop, Disk Bytes per Flop)

Figure 13. NERSC's science-driven systems in 2009.

2010

- NERSC 6 in full service.
- Upgrade WAN link to 100 Gb/s.
- Storage and Facility-Wide File System upgrade.

5. Science-Driven Services

The goal of NERSC's Science-Driven Services is to make it easy, practical, and productive for all DOE computational scientists to use the NERSC high-end systems. NERSC will work with its users to help lower the gap between the peak performance of terascale systems and the performance realized by applications running on them.

NERSC continues to affirm its commitment to excellent technical support for its user community of about 300 projects and more than 2500 users. It is the heart of the strategy that sets NERSC apart from other sites and greatly enhances the impact of NERSC's high-end systems. Users consistently give NERSC's support services high ratings in the annual user surveys.

NERSC provides Science-Driven Services by:

- Constantly soliciting feedback from users via monthly teleconferences with the NERSC User Group and annual User Surveys, as well as through the consulting process
- Understanding and addressing the unique issues of using large-scale systems
- Providing consistent, high-quality user support (direct scientific support)
- Collaborating with scientists on major projects that require extensive scientific computing capabilities (collaborative scientific team support)
- Ensuring that the production systems and services are the highest quality, stable, secure, and replaceable within the constraints of budget and technology (system monitoring and management)
- Aggressively incorporating new technology into the production NERSC facility by working with other organizations, vendors, and contractors to develop, test, install, document, and support new hardware and software (system improvements)
- Maintaining open, unencumbered access at high performance rates so that NERSC systems can effectively be used for science while simultaneously minimizing disruptions due to computer security attacks (security)

5.1 Direct Scientific Support

Direct scientific support focuses on helping the DOE scientific community become more productive in its computational and data management work. It is important that the user community be able to ask for assistance in the way most effective for them — not just what is most efficient for NERSC. Thus, NERSC supports telephone, email, and Web interactions with timely acknowledgement and response resolution and escalation. Once a user reports a problem, NERSC manages it until it is resolved, not just sending the user to another group or having the

user manage the problem. NERSC Operations provides basic user support around the clock, including password change requests and management of system problems, with live consulting phone coverage from 8 a.m. until 5 p.m. (local time) Monday through Friday.

The key components of this activity are described below.

Science-Driven Consulting

NERSC consultants are HPC experts with advanced degrees in astrophysics, chemistry, computer science, physics, mathematics, and statistics. They resolve user problem reports and requests for assistance, particularly with regard to programming and application development. They help analyze and debug problems with user codes, as well as with systems and applications software; report problems to vendors; track problems so they will be corrected in a timely manner; and introduce new systems and technologies to the user community.

Typical issues that are handled by the consultants include:

- analysis of parallel performance and scaling optimization
- code performance tuning
- efficient use of math, graphics, and message passing libraries
- algorithmic code restructuring to increase performance
- debugging code that causes error conditions
- I/O optimization and parallel I/O strategies
- effective use of compilers
- strategies for porting code.

Since the consultants work closely with NERSC users, they can help assess user needs and requests and advocate for them to the NERSC organization as well as the vendor community. The consultants provide input on how best to configure center resources to meet user needs. With flexible, well-trained staff, NERSC is well poised to meet upcoming challenges:

- new algorithms and optimization techniques
- new languages
- increasing parallelism
- emerging HPC systems technology
  - hybrid processor systems
  - parallel file systems, parallel I/O profiling

Applications Software Support

Consulting staff install and provide software support for a complex set of applications, libraries, tools, and environments, such as NAG, MPI, Gaussian, TotalView, and performance analysis tools. Over 200 different software packages are supported, some of them from the open scientific community.

The overall goal of NERSC's software management process is to ensure that users have easy access to up-to-date software to maximize their productivity and research capacity. The consulting staff keep abreast of developments in software available for high performance computers. They monitor developments in vendor-provided software and select new offerings to bring to NERSC. In addition, NERSC users may submit requests for new software.

Consulting staff manage application software bugs, reporting the problems to the vendors, interacting with the vendors to provide as timely a solution as possible (including escalating the problem to the vendor), and often writing code to demonstrate the problem to the vendor. In the process of installing software from the open scientific community, problems are often encountered, analysis is performed to understand the problem and resolve it, and the resolution is sent back to the software maintainers to be incorporated in future releases. Thus NERSC staff play an important role in improving the quality of scientific software.

The consulting staff are also intimately involved in testing and evaluating the changes brought on by new system functions. They assist in developing and maintaining test software suites that are used in system regression testing, and they test new application software such as compilers, libraries, and tools for advanced programming. New software goes through a pilot stage (staff testing followed by user testing) before being installed as production software.

Consultants also write tools for users to enhance their productivity. An example is the Integrated Performance Monitoring (IPM) tool that was developed to help users identify scaling bottlenecks in their codes. IPM is easy to use, has low overhead, and scales to thousands of processors. Sample output from IPM is shown in Figure 14.

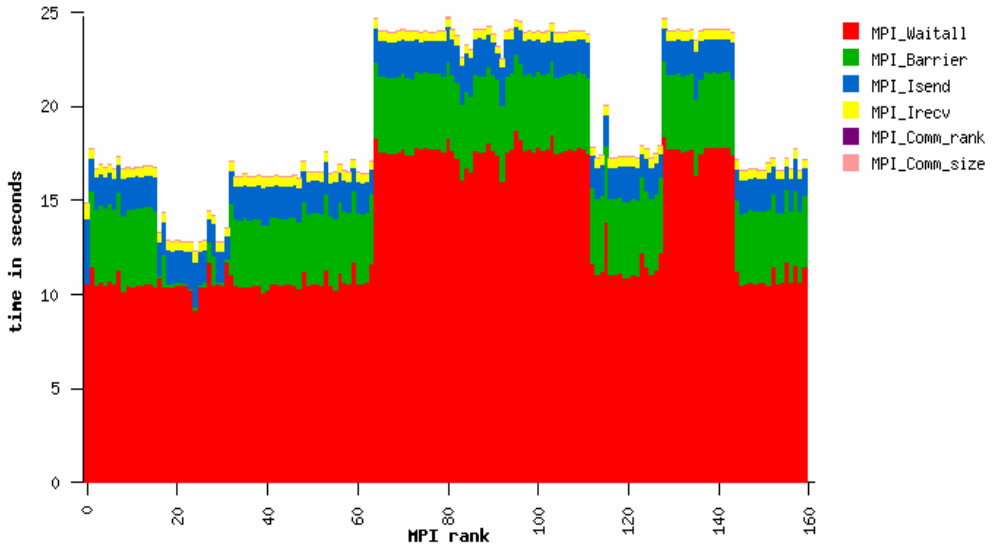


Figure 14. IPM can graphically present a wide range of data, including communication balance by task, sorted by MPI rank.

Web Information, Documentation, and Training

The consulting staff maintains the NERSC Web site at www.nersc.gov. The consultants write and maintain user documentation with the goal of providing timely and accurate information for all systems and software. Additionally, they organize documentation provided by vendors, and often augment and summarize the vendor documentation.

The NERSC Web site displays "active information" such as machine statuses, job statuses, and usage statistics. It also provides customized job performance information. Users can drill down from active queue displays to get job and MPI performance information for their jobs, as shown in Figure 15.

Function	Total calls	Total time (sec)		Total buffer size (MB)	Avg. Buffer Size/call (Bytes)
MPI_Barrier	6.02e+05	3.48e+05	44.23%	0	0
MPI_Allreduce	3.18e+07	2.31e+05	29.33%	3.61e+05	11,936
MPI_Send	1.29e+08	1.29e+05	16.36%	5.24e+04	426
MPI_Bcast	5.73e+07	6.08e+04	7.73%	5.39e+04	987
MPI_Reduce	1.08e+08	1.24e+04	1.58%	1.66e+05	1,620
MPI_Recv	1.29e+08	6.11e+03	0.78%	5.24e+04	426
MPI_Comm_rank	1.14e+03	5.92e-01	7.52e-05%	0	0
MPI_Comm_size	6.66e+02	0	0%	0	0

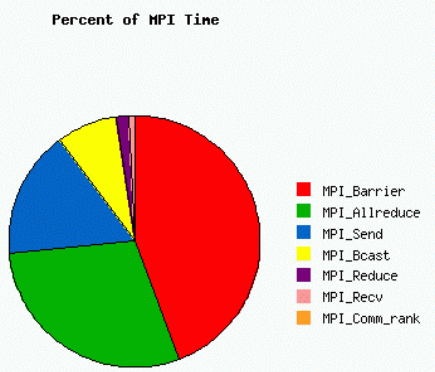


Figure 15. An example of job performance information available on the NERSC Web site.

NERSC provides advanced training instruction in the use of the latest technology. These activities include multiple days of intensive classes, lectures, seminars, and symposia presented in collaboration with other groups. In addition, many users receive training through individual consultations, and by reading the user documentation on the Web site, especially the online tutorials. User comments indicate that many are able to train themselves from NERSC's Web pages:

- "Very useful, and I was able to get up and running reading them, without needing much other help."
- "It's one of the best places to find information about HPC."
- "NERSC Web pages are really, really good and well organized."

Account Management and Allocations Support

The accounts and allocation support staff maintain and extend several major tools that provide NERSC users with the ability to manage their project resources. The core part of the effort is in the use, support, and extension of the NERSC Information Management system (NIM). This system manages all accounts and projects for NERSC systems, automatically installs accounts, accumulates and summarizes usage data of users and projects, and implements resource restriction if a project or user exceeds the allocation. NIM also provides Principal Investigators and DOE Program Managers with the ability to manage their project resources. All accounts and allocations are periodically reviewed and validated, with unused accounts being disabled.

NERSC manages the allocation request process, using the part of NIM called the Energy Research Computing Allocations Process (ERCAP). ERCAP includes an interface for reviewers to review and score the requests and for awards to be assigned to the requests. The DOE Office of Science makes award decisions for production-level requests, while NERSC makes decisions for Startup requests.

5.2 COLLABORATIVE SCIENTIFIC TEAM SUPPORT

NERSC works directly with scientists on major projects that require extensive scientific computing capabilities. These projects are characterized by large collaborations, the development of community codes, and the involvement of computer scientists and applied mathematicians. In addition to high-end computing, these large projects handle issues in data management, data analysis, and data visualization, as well as automation features for resource management. The goal is to collaborate intensively with six to twelve large projects every year. NERSC expects that new services will be created as a result of working closely with these projects, and that some of these services will then become part of the NERSC environment, benefiting the larger NERSC user community. Examples include the CVS server, which was originally deployed to assist a SciDAC project and is now available to all NERSC users, and the remote visualization license server, which was originally deployed to facilitate the local visualization needs of an INCITE project.

NERSC takes a multifaceted approach to its support of high-end teams. Computational and/or disciplinary experts are assigned to each team as a point of contact (POC). The POCs not only directly collaborate with the project, but also serve as an advocate for the project within NERSC. Part of the POC's job is to coordinate with other NERSC staff to resolve issues and requirements the teams have, and to coordinate services provided to the team. These services include:

- intensive code tuning, scaling and optimization
- additional system resources (e.g., TB of disk, job scheduling boost)
- help incorporating or developing new algorithms
- help creating a data management, analysis, and visualization framework
- developing an end-to-end computing environment that integrates experimental facilities, computing, storage, and analysis.

Some projects will need to integrate high-end computing, storage, and data management capabilities into a distributed scientific experiment environment. A compelling reason for the integration of simulation with experiment is that each can drive and modify the other in a time-constrained interaction. Data-driven projects like supernova cosmology that use observation to refine simulation, and simulation to drive observation within a narrow time window, and use large-scale data archives to provide persistence for both, cannot be done without this sort of integration. Grids and Web portals are examples of middleware that can tie together the computational simulation and data management environments and facilitate building multi-component applications that execute codes; catalogue, store, and access data; and integrate collaborators at different locations. NERSC will be involved in facilitating and sometimes developing these end-to-end environments for large projects.

In addition, NERSC supports many application scientists by collaborating directly with them as peers on their projects. Community codes are developed as part of these collaborations; examples include a fully-coupled global climate model; a parallel I/O library for efficient parallel data I/O in climate models; a fusion code for the study of long-wavelength, low-frequency, nonlinear phenomena in realistic toroidal geometry; efficient computational methods for excited state calculations, calculations of electron transport, and electronic structure calculations based on scattering theory; a parallel plane wave pseudopotential program for atomistic total energy calculation based on density functional theory; noise estimation for cosmic microwave background data analysis; and new eigensolvers. The results of these collaborations are documented in joint publications or presentations at scientific conferences.

5.3 COMPUTING INFRASTRUCTURE SUPPORT

NERSC deploys the best and most appropriate technology in a cost-effective manner, and continuously monitors and tunes computing systems and software to ensure security, stability, and high performance.

System and Network Monitoring and Support

NERSC staff monitor all aspects of the NERSC Center on a 24 × 7 × 365 schedule. These tasks involve system and network monitoring, initial system troubleshooting, system backup, and management of the near-line and off-line storage media. NERSC uses productivity-improving tools and techniques designed to automate or improve these operational tasks.

NERSC systems and storage staff provide basic system administration and remedial maintenance around the clock. Each system has an assigned point of contact who responds to system issues and problems. A lead system manager is also assigned to each system, who is responsible for overall operation and support as well as being an expert in the particular hardware and software. Vendor personnel, sometimes on site, are available to ensure that the systems operate well and provide high reliability.

System Management

System management is the term used to describe system administration as well as the advanced tasks of resource management, system tuning, system improvement, and developing new functionality. For example, by aggressively using advanced scheduler functions, NERSC has been able to double the amount of computational capability delivered by some systems, compared to that delivered by the standard vendor software. Systems have complex interactions of memory, CPU time, I/O, and networking that make it challenging to balance high utilization and fast turnaround for a diverse set of users and disciplines. NERSC constantly performs tuning and balancing to assure that the resources are well used but also have the best possible response time.

NERSC responds quickly to special requests from the NERSC community for processing and services. NERSC is able to provide highly effective processing — be it high priority, very long runs, massive data, or other special needs — because the system managers can configure the systems to respond to multiple needs for resources.

Proper system management includes cyber security. NERSC uses a state-of-the-art approach, applied with the philosophy of ensuring that known security problems are fixed, systems and communications are monitored for inappropriate activity, and security incidents are responded to swiftly. NERSC uses and improves advanced monitoring and reactive tools that limit inappropriate access but provide the best level of security with minimal impact on performance and function.

System Improvements

NERSC system staff members are highly skilled at testing and integrating new system hardware and software with little or no service disruption. Technology comes from vendors, the open-source community, the academic community, national laboratories, NERSC staff, and other sources. NERSC's job is to deploy the best and most appropriate technology in a cost-effective manner. Technology — whether hardware, software, or a combination — enters the process and goes through different phases based on its source, maturity, and function. The phases are:

- experimentation and development
- evaluation, observation, and external testing
- testing
- early use
- general or special use
- full service.

At each phase, the technology is evaluated for:

- readiness to progress to the next phase
- potential impact on NERSC clients (both immediate and long term)
- overlap with existing functions
- costs (both initial and ongoing)
- benefits and risks.

NERSC reviews the status of the technology before it is allowed to progress to the next phase. Throughout the phases of the process, NERSC staff provide feedback and analysis of the technology to the supplier. NERSC may assist in the development of new requirements for vendors and other groups, and, when appropriate, develop key technology that is important to the success of NERSC clients. NERSC staff interface with vendors and with other sites and visitors in this process.

System Software Infrastructure

In order to deliver the best possible environment to users, NERSC develops and deploys software that is not associated with any particular computational systems. Some examples of such infrastructure include:

- *NERSC Information Management (NIM).* NIM provides centralized account management for NERSC users. NIM manages the entire project lifecycle, allowing principal investigators to

submit proposals, manage accounts, view allocations and usage, and a number of other functions. NIM was developed at NERSC.

- *LDAP single sign-on.* NERSC has architected and deployed a single sign-on mechanism based on LDAP that allows users to have a single password on all production machines, and provides administrators a consistent authentication and authorization interface.
- *One-time passwords.* NERSC is a leading participant in an effort by DOE labs to design and implement an infrastructure that supports one-time passwords (also known as two-factor authentication) for all access to NERSC systems.
- *System monitoring.* NERSC configures and modifies third-party tools and when necessary develops its own tools to provide comprehensive system monitoring, including security monitoring.
- *Workload profiling.* NERSC has developed the Integrated Performance Monitoring (IPM) tool (described in Section 5.2 above) which makes application profiling data available to users. IPM also provides information to NERSC staff, enabling NERSC to better understand the requirements of its workload.

Software infrastructure development and deployment will continue to be an important part of providing a high-quality environment to NERSC users. One long-term trend driving the increased importance of this activity is increasing reliance on horizontal integration and open-source software.

Security

NERSC resources are accessed via ESnet to fulfill NERSC's mission as a national user facility. Even though ESnet is part of the Internet, NERSC has been remarkably successful in maintaining open, unencumbered access at high performance rates so that NERSC systems can effectively be used for science while simultaneously minimizing disruptions due to computer security attacks. For example, during the spring 2004 computer security incident when over 2000 university and government systems were seriously compromised and many computer centers were shut down for weeks, NERSC was one of the few major computing centers that did not have to shut down its high performance computers, even though we were attacked multiple times.

NERSC has been very successful in implementing effective computer security measures while maintaining the unencumbered ability of NERSC users to do science. NERSC plans to continue using the following:

- a limited border firewall using access control lists (ACLs) that blocks unsafe protocols such as telnet without incurring the performance degradation or default deny policy of traditional firewalls
- border monitoring and intrusion detection systems such as Bro that can automatically detect and block attackers
- border recording of network traffic for later forensic analysis
- network subnets based upon function, with ACLs in place to limit network exposure — for example, all NERSC Web servers are located on one subnet, and Web server protocols are restricted to all other subnets

- internal network monitoring and intrusion detection systems
- a protected, central audit server that collects system logging data from all systems and netflow/sflow data from network equipment
- internal firewalls to protect specialized subnets (e.g., system control workstations)
- vulnerability scanning to detect misconfigured or unpatched systems
- aggressive patch management to fix vulnerabilities
- firewall software on individual systems
- system monitoring and intrusion detection
- virus and spam filtering on all incoming email
- antivirus software for Windows systems.

Each year, the number of computer security attacks increases, and they are becoming more sophisticated. A few years ago, our primary security concern was attacks on vulnerable system services. Today the biggest security issue is authentication credential theft. Tools are readily available that allow an attacker to sniff user passwords and log on to systems as a legitimate user. One way of mitigating this attack is to install a one-time password (OTP) system. NERSC has started an OTP implementation using hardware tokens. The implementation is being carefully planned to minimize negative impacts to NERSC users. Unfortunately, OTP is not a panacea for credential theft. As OTP systems become more widely implemented, NERSC expects attackers to rapidly develop tools that automatically hijack a valid user session. The attacker will let a user log on to the system using their OTP hardware token and then take over their connection once authentication is completed. New computer security techniques are required to counter this and other attacks.

NERSC will continue to analyze and test commercial and non-commercial computer security hardware and software as well as enhance new security software. Areas of particular interest are:

- better and more comprehensive internal intrusion detection systems
- integrating network and host security data within NERSC
- cross-site integration of network and host security data
- additional monitoring of new and experimental protocols
- improved user activity verification on systems
- central auditing of ssh session content
- automated tools for security log analysis
- automated compliance tools.

As more and more science depends on access to high performance computing resources, the role of computer security becomes more important. Since much of computer security is labor intensive, the amount of NERSC staff time required for computer security will increase. Unfortunately, another effect of increasing computer security attacks is an increase in government compliance requirements, audits, and restrictions. Computer security has become

bureaucracy intensive, and NERSC will be required to devote more staff to compliance as well as technical issues.

Server Support

NERSC staff will continue to build, test, and support general servers and software that support NERSC users and staff. Examples include the NERSC Information Management (NIM) system, email, Web, LDAP, software licensing, and CVS servers. Many of these systems are supported 24 × 7 × 365.

NERSC routinely evaluates software that could increase the productivity of our users and staff and welcomes such requests from NERSC users. The CVS source code repository, which was initially implemented for SciDAC users, was later made available to all NERSC users. Much software today is open-source software; while such software can be production-ready, it frequently comes with minimal or no options for support. In such cases, NERSC staff must provide in-house support.

5.4 OUTREACH

NERSC will engage in outreach activities to

- promote NERSC as the premier high performance computing center
- help train future computational scientists
- attract new users to the NERSC facilities.

These activities will include the following:

- *Startup Program.* NERSC will continue the Startup Program to attract new users to NERSC. Each startup allocation ranges from 10,000 to 20,000 node-hours. On average, the Startup Program represents about one percent of the entire annual usage. The intent is to allow a new user to gain experience using high performance computing with the hope that, if successful, the user will apply for a regular allocation and become a long-term NERSC user. Every year about one-third of the Startup projects make the transition to production status, one-third remain in the Startup program for a second year, and one-third do not continue at NERSC.
- *NERSC Web site and NERSC News.* The NERSC Web site and the bimonthly newsletter "NERSC News" provide mechanisms to report and promote computational work performed at the NERSC Center, including major scientific accomplishments that have been enabled by high performance computing. The Web site and newsletters also serve as a forum to present to the NERSC users and others information on tools and resources for high performance computing.
- *Participation in selected conferences and workshops on high performance computing and computational sciences.* Staff represent NERSC at conferences and workshops such as the SciDAC PI annual conferences, the Computational Engineering and Sciences Conferences (CESC), and SciCOMP, the IBM Scientific Computing User Group.
- *High Performance Computing Lectures.* Lectures provide a mechanism to educate and train future parallel programmers and computational scientists. The annual DOE Advanced CompuTational Software (ACTS) Collection workshops provide hands-on instruction in

building robust scientific and engineering high-end computing applications. Additional funding might be secured to hold High Performance Computing Summer Institutes.

5.5 READINESS FOR NEW CHALLENGES

We believe that our existing service model will continue to serve NERSC users into the future. With well trained and flexible staff, NERSC is well poised to deploy staff resources to meet upcoming challenges in:

- new algorithms and optimization techniques
- new languages
- increasing parallelism
- emerging HPC systems technology
- analytics.

There will be increased focus on user training, which will be critical in introducing new technologies to the NERSC user community. There will be an increased focus on intensive collaborations with projects in order to support their scientific needs to collect, manage, and analyze data and to create end-to-end computing environments. Finally, there will be more partnerships with the larger HPC community to create science-driven computer architectures and to foster new HPC ecosystems.

6. Science-Driven Analytics

The term *analytics* refers to an emerging set of interrelated technologies that combine to produce insight and understanding from large, complex, and disparate datasets. More specifically, the term *visual analytics* is an "outgrowth of the fields of scientific and information visualization but includes technologies from many other fields, including knowledge management, statistical analysis, cognitive science, decision science and more. The processes and goals of analysis dominate the approach, but it's enabled by the wide-band visual interface to the brain and a dynamic interaction style of communication and discourse.[6]" Visual analytics is the science of analytic reasoning facilitated by interactive visual interfaces.[7] Its objective is to enable analysis of overwhelming amounts of disparate, conflicting, and dynamic information, and requires human judgment to make the best possible evaluation of incomplete, inconsistent, and potentially erroneous information.

The science of analytics brings many disparate technologies to bear on a relatively simple idea: understanding data. The technologies include (1) data management, (2) visualization, (3) analysis, and (4) discourse. These in turn rely on the fabric of computational infrastructure, expertise in applying methods on such infrastructure, and close cooperative interaction between domain scientists, computational scientists, and computer scientists.

---

[6] P. Wong and J. Thomas, "Visual Analytics," IEEE Computer Graphics and Applications **24** (5), Sept/Oct 2004.
[7] J. Thomas, ed., "National Visual Analytics Research Agenda," http://nvac.pnl.gov/research.stm.

1. In broad terms, *data management* refers to storage and retrieval of scientific data from primary, secondary, tertiary, and distributed/federated sources; data description, organization, and metadata management; efficient queries; format integration; distributed data management and movement across networks; data storage and caching.
2. *Visualization* refers to transforming abstract data into images, and relies on high-bandwidth human cognitive processing to produce understanding.
3. *Analysis* describes activities that aim to derive, compute, or locate features or characteristics within data. The distinction between visualization and analysis becomes blurry where the fields overlap.
4. *Discourse,* which is a relatively new term in this context, refers to the process of analytic reasoning or knowledge discovery. Discourse is interaction leveraging the underlying technologies — data management, visualization, and analysis — but aimed at producing specific types of understanding. Discourse, which is the expression and practice of analytics, is arguably one of the most challenging technical aspects of analytics.

6.1 THE ROLE OF ANALYTICS IN SCIENTIFIC RESEARCH

History has shown that scientific research benefits directly from advances in technology. The technology advances, which range from improvements in computing, storage, and networking to more sensitive instruments, make it possible to create, gather, and store vast amounts of data. The trend is towards science that is dominated by information analysis and understanding. We are quite literally producing and storing data at a rate faster than we can understand it.

No single "analytics pipeline" or visual presentation paradigm can address all possible tasks and situations. Some tasks are time-critical, requiring data to be analyzed and results produced within a given period of time to take advantage of a window of scientific opportunity. Examples abound, from the need to modify apparatus or instruments between experimental runs, or to converge a community of telescopes on a short-lived astral event upon discovery. Typically, tasks of this sort can be characterized as "confirming the expected," or more generally, proving the presence or absence of a known phenomena. Conversely, analytics also plays a key role in discovering the unexpected, which is a hallmark of scientific discovery.

It is generally accepted that analytics implementations are a combination of technologies from diverse origins, ranging from research prototypes to commercial, off-the-shelf tools. Analytics implementations are an integration and deployment of many different types of technologies. The most effective analytics solutions tend to be those tailored to a specific application science domain to perform a specific task or family of related tasks.

An example of time-critical analytics performed on experimental data is the work of the Nearby Supernova Factory (SNfactory).[8] Discovering supernovae as soon as possible after they explode requires imaging the night sky repeatedly, returning to the same fields every few nights and quickly post-processing the data. The most powerful imager for this purpose is the charge-coupled device (CCD) camera built by the Jet Propulsion Laboratory. It delivers 100 MB of image data every 60 seconds, and an upgraded version of the camera will more than double its

[8] http://snfactory.lbl.gov/

resolution. The new images are computationally compared to previous images of the same field using digital image subtraction to find the light of any new supernovae. Because the amount of data is so large (50 GB per night per observatory, or 18.6 TB per year), the image archive even larger, and the computations so extensive, it is critical that the imaging data be transferred to a large computing center (in this case NERSC) and processed as quickly as possible. The refined data is then analyzed and compared to theoretical simulations in order to select candidate stars to watch more closely. The candidate list is then distributed to observatories around the world. The workflow aims to detect and report new supernovae targets with less than a 24-hour turnaround so that researchers around the world may train their telescopes on the event, which is relatively short lived. The distributed analytics workflow has had a profound and positive impact on cosmological research. Prior to the SNfactory distributed analytics workflow, around 130 Type Ia supernovae had been discovered. Since its deployment, the SNfactory workflow discovers new Type Ia supernovae at a rate of about 8 to 9 per month. This project is also contributing to the design of future experiments, such as the Supernova/Acceleration Probe (SNAP),[9] a satellite that is now being developed.

The SNfactory is itself an offshoot of the Supernova Cosmology Project,[10] which provided a prime example of analytics discovering the unexpected — in this case, the accelerating expansion of the Universe, one of the key scientific discoveries in recent times. To analyze their data from 40 Type Ia supernovae for errors or biases, the Supernova Cosmology team used a NERSC supercomputer to simulate 10,000 exploding supernovae at varying distances, given universes based on different assumptions about cosmological values; these were then plotted and compared with real data to detect any biases affecting observation or interpretation. The unexpected conclusion of this analysis was named *Science* magazine's "Breakthrough of the Year" for 1998.

6.2 NERSC'S ANALYTICS STRATEGY

NERSC's analytics strategy builds on two of the Center's existing strengths: (1) proven expertise in effectively managing large, complex computer systems, infrastructure, and datasets to solve scientific problems of scale, and (2) exemplary user services, consulting, and domain scientific knowledge that help the NERSC user community effectively employ the Center's resources to solve challenging scientific problems. On this foundation, NERSC's analytics strategy adds an increased emphasis on facilities, infrastructure, expertise, and alliances that provide or employ the technologies used to realize analytics solutions. NERSC will also leverage Berkeley Lab's world-class scientific data management efforts.

NERSC is realigning its current resources to support analytics activities. The notion of "Center infrastructure" is being broadened to include elements such as Web services, database deployment and support, and so forth. An increased focus on data analysis and scientific data management will serve to support analytics. The existing visualization program will be expanded to include a broader charge that includes information visualization and integrated data management, analysis, and distributed computing. The goal is a well-rounded service and

[9] http://snap.lbl.gov/
[10] http://www.lbl.gov/supernova/

technology portfolio that is responsive to the analytics needs of NERSC's user community. Several aspects of NERSC's analytics strategy are discussed in more detail below:

- taking a proactive role in deploying emerging technologies
- enhancing NERSC's data management infrastructure
- expanding NERSC's visualization and analysis capabilities
- enhancing NERSC's distributed computing infrastructure
- understanding the analytics needs of the user community.

Taking a Proactive Role in Deploying Emerging Technologies

The majority of software deployed by NERSC can be characterized as either commercial or production-quality software. Research-grade software is occasionally deployed, but that is the exception rather than the rule. The proliferation of open-source software over the past few years has taught us that "free" does not mean "zero cost." There is ample anecdotal evidence that "free software" can incur a substantial net cost to the Center, particularly if it introduces new security vulnerabilities or otherwise requires significant staff time to tune and adapt it for use in the NERSC environment. To maintain high productivity, NERSC prefers commercial or production-quality software over research-grade software.

On the other hand, NERSC does have a history of success in being an early adopter of new technologies and in working closely with developers to help technologies mature, when those technologies offer a substantial and immediate benefit to our users. For example, our work with the HPSS consortium over the years has had a positive impact in helping to harden the HPSS technology as well as providing a robust storage solution to the NERSC user community.

With an increasing emphasis on analytics, there will be a corresponding increase in the role played by research-grade software in analytics solutions that are deployed in domain-specific workflows. In its support for analytics activities, NERSC will increasingly become a conduit for prototype technologies that emerge from the DOE computer science (CS) research community. That community is directly responsible for advances in fundamental technology areas germane to analytics: data management, visualization, analysis, discourse, computer networking, programming languages, component interfaces, frameworks, and so forth. Realizing specific analytics workflows, like the Supernova Factory example above, will require adapting and deploying technologies from several different areas — data management, analysis, visualization, dissemination — into a unified analytics workflow that functions effectively in a production environment and possibly at a time-critical pace.

The process of transitioning software from diverse origins and purposes into broad analytical use is complex, and it requires the cooperative efforts of researchers, software engineers, systems infrastructure and operations staff, training and support staff, and the users themselves. The role of NERSC staff will include deploying new system and support software, helping applications software engineers effectively use center resources, and playing a proactive role in providing feedback to the original CS researchers and developers to address security or performance concerns. Given resource constraints, NERSC will focus on those technologies and

collaborations that promise to have the broadest possible positive impact across its user community.

Enhancing NERSC's Data Management Infrastructure

A significant challenge facing all sciences is information management and understanding. Scientific data management has been identified as a bottleneck in many fields of scientific endeavor,[11] and new data management technologies will play a pivotal role in scientific success. NERSC's scientific data management strategy is a multifaceted approach that focuses on several interrelated topics. One is the ability to store and retrieve more data, and do so more quickly. Another is the ability to more quickly find data of interest in large, complex datasets. A third is the ability to share effectively share data among distributed communities of scientific researchers.

In 2005–2006 the NERSC Center plans to deploy a Facility-Wide File System (FWFS). The FWFS will offer increased performance for all applications, including data-intensive analytics tasks. It also will help streamline workflows in that files no longer need to be copied from resource to resource. It is intended to provide extremely high I/O rates, which will be of increasing importance with increasing data sizes. Another dimension of NERSC's data management infrastructure is the planned increase in archival storage to nearly 40 PB over the next five years, as described above in Section 4.5. Together, the increase in archival storage and the Facility-Wide File System will help serve the data management infrastructure needs of the user community.

A cornerstone of analytics is being able to find data that matches specific search criteria. In some instances, commercial or standard database technology is appropriate. In other instances, the best solution is a research product from the CS research community that provides capabilities unmatched in commercial or freeware sources. NERSC will track and deploy appropriate data management software used for efficient application-level data storage and retrieval. While NERSC can be expected to provide frontline support for standard database interfaces (e.g., SQL), NERSC staff are not experts in the design of databases, nor are they currently well versed in design or use of research-grade index/query technology. As the Center grows its program to support analytics, we anticipate an overall growth in the knowledge and experience base in these areas.

Distributed data management is another area that will receive more emphasis in the future. NERSC has a successful track record in this area already: the center helped facilitate deployment of logistical networking to facilitate movement of fusion simulation data from NERSC to Princeton Plasma Physics Laboratory for storage and subsequent analysis. As community data collections become more prevalent, NERSC will track and deploy appropriate technology in response to direct requests from its user community. One potential technology for future deployment is the Storage Resources Manager (SRM) developed by the Scientific Data Management Group at Berkeley Lab. SRMs provide dynamic space allocation and file

[11] In 2004, a series of scientific data management workshops was hosted by Stanford Linear Accelerator Center to better understand the scientific data management needs of the application sciences. The report for the workshop is available at http://www-user.slac.stanford.edu/rmount/dm-workshop-04/Final-report.pdf.

management on storage resources shared among distributed computing resources. SRMs are used as part of the Earth System Grid[12] to provide distributed, file-level data management.

Expanding NERSC's Visualization and Analysis Capabilities

Visualization and analysis are the most visible elements of analytics. The NERSC Visualization Group's mission historically has been to deploy and apply visualization technology to support the science needs of the computational science community hosted at NERSC. One of the most significant activities performed by the visualization staff is in-depth, one-on-one consulting services, such as those provided to two of the 2004 INCITE projects (Figures 16 and 17). These activities typically involve finding or engineering solutions where none exist in off-the-shelf form. In addition, visualization staff evaluate new visualization hardware and software technologies to determine which are beneficial to the Center and to the user community.

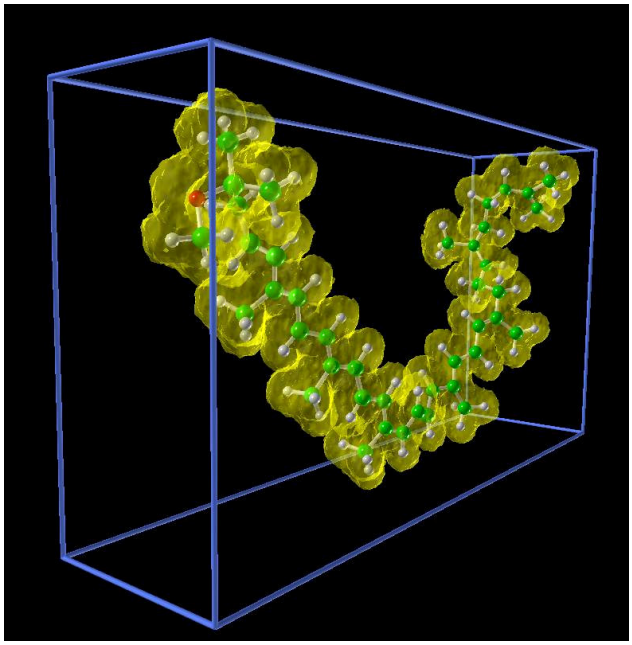


Figure 16. Electron density of the spheroidene molecule, a regulatory agent that dissipates excess energy, thus preventing oxidation damage during photosynthesis.

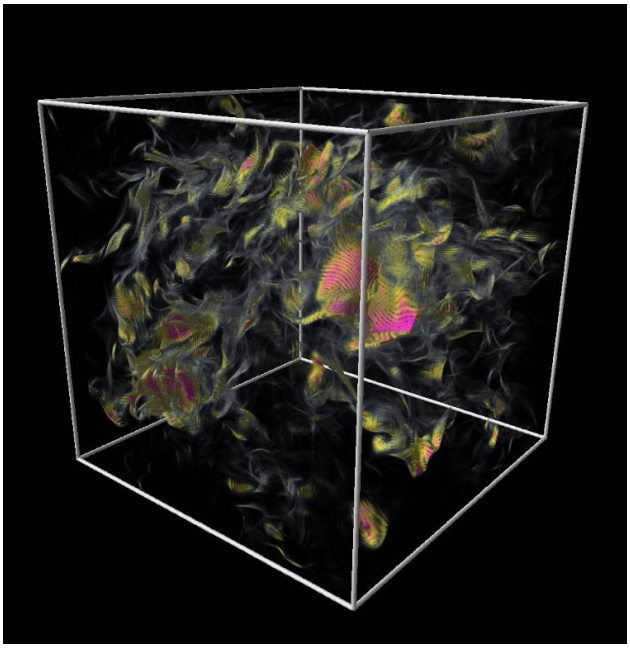


Figure 17. Data values from a 3D direct numerical simulation of turbulence at high Reynolds number were "volume rendered," i.e., mapped onto color and opacity, to reveal features of the data.

The visualization needs of NERSC users are diverse, and range from the need for scalable, high capability visualization to data management standards that form the basis for entire computational science projects. The Visualization Group is addressing the following user requirements:

- visualization and analysis tools (and facilities) capable of processing large scientific datasets
- the need to leverage parallel computing resources for data-intensive analysis and visualization
- tools and techniques to effectively deliver visualization technology to the NERSC user community

[12] https://www.earthsystemgrid.org/

- visualization tools that support interactive, multivariate exploration, including "drill-down" capabilities as well as the means to easily transition between macro and micro views
- centralized management and deployment of visualization technology
- tools and facilities for automated (and offline) data analysis, visualization, and exploration
- better integration between data management, data analysis, and visualization.

NERSC's forte over the years has been scientific visualization, with particular emphasis on high performance, parallel scientific visualization to effectively support its remote user community. A less mature dimension of the field is known as *information visualization.* (By "less mature," we mean that there are far fewer ready-made applications for information visualization.) Information visualization differs from the better-known scientific visualization in that the underlying data does not readily lend itself to spatial mapping. The familiar visualizations of climate simulations are an example of scientific visualization — water vapor, temperature, or some other variables are mapped onto a projection of the earth's surface. In contrast, many applications do not enjoy such ready-made spatial mapping. For example, there is no natural and intuitive spatial mapping to compare the results of genome alignment across multiple species. These problems and challenges are exacerbated by high-dimensional and time-varying data.

Analysis plays a central role in analytics, for it offers the means to identify and track features in data over time. The means by which features are identified and tracked are the subject of much current analysis research. As data size and complexity grow, it will become increasingly crucial to use analysis technologies to reduce processing load through the computational and visualization pipelines, as well as to reduce the "scientific processing load" on the human who must interpret and understand the results. Techniques abound for performing just these activities: principal or independent component analysis, spectral analysis, and so forth.

For both visualization and analysis, the overall NERSC strategy will be first to understand the needs of its user community. With a clear understanding of needs and priorities, NERSC will be in a position to gauge what technologies it will provide to have broad impact in meeting the analytics needs of its user community. NERSC will continue to use its successful model of team-centered support to bring appropriate technology to bear on analytics problems. As with the other technologies that comprise analytics, we anticipate that a portfolio of commercial, production, open-source, and research-grade technologies will be most effective. Our vision is to expand our basic service concept more broadly into the field of analytics — to include information visualization, analysis, data management, discourse, and effective use of center facilities. We have had early success with this type of model in the comprehensive team support afforded to the INCITE projects in 2004.

Enhancing NERSC's Distributed Computing Infrastructure

The SNfactory project is but one of many that illustrate the need to combine distributed computing and data resources into a single analytics workflow. Because there are so many permutations of technology and location, NERSC's strategy for supporting distributed computing will be tailored to provide service that has the broadest possible benefit and that conforms to security and related requirements. While there is relatively little risk in deploying a research prototype visualization application, there is much greater risk in deploying a new form

of authentication that may expose the Center to a security incident that could result in loss of service, data, or even property.

NERSC will provide the low-level infrastructure that is commonly used for distributed computing services. Such infrastructure includes the Open Grid Services Architecture (OGSA) and similar Grid-related technologies that provide authentication and secure data movement across the network. Higher-level applications and services that emerge from the research and applications communities should be engineered to be portable across authentication and transmission modalities to have the maximum portability and usefulness to the user community. Much prior work in this space has been performed by projects such as the Particle Physics Data Grid[13] and the Earth System Grid[14]. Both of those projects rely on standard services for brokering access to data and tools that serve large, distributed user communities. NERSC will be proactive in tracking and deploying such technology, and will work closely with the user community to provide them the documentation and assistance they need to construct analytics workflows. An early success in such activities was our deployment of logistical networking to aid Princeton Plasma Physics Laboratory researchers in staging data from NERSC for movement to Princeton, where it is cached and shared by a team of researchers. Similar opportunities may arise where a small amount of extra effort on the part of the NERSC staff will result in new functionality that can be leveraged by multiple user projects.

Understanding the Analytics Needs of the User Community

Over the years, the NERSC user community has proactively provided guidance and input to the program in the form of the "NERSC Greenbook." The Greenbook has generally consisted of individual sections describing various fields of computational science and the types of computing resources they require. It is generated on a three-year cycle that roughly corresponds with the procurement of each new computational platform. NERSC also conducts an annual user survey aimed at evaluating user satisfaction and discovering user "hot spots" within the program. Since the Greenbook and annual user survey tend to focus on computational and storage requirements more than visualization or data analysis, in 2002 NERSC hosted its first "Visualization Greenbook" workshop.[15] The findings document contains detailed descriptions of visualization and analysis needs, and has been instrumental in shaping and prioritizing the existing NERSC visualization effort. To be effective, NERSC's new program focus on science-driven analytics will require additional information from the user community in the areas that include analytics: data management, visualization, data analysis, discourse, distributed computing, networking, and so forth.

7. Investment Strategy and Budget

NERSC has an investment strategy that will enable us to complete this plan. NERSC will continue to invest in computational systems and system balance in the same proportion as in the past five years. NERSC is reducing staff by about 10% to offset increases in electricity, maintenance, and other operating costs. This decrease was done during 2004–2005 through

[13] http://www.ppdg.net/
[14] http://www.earthsystemgrid.org/
[15] http://vis.lbl.gov/Events/VisGreenbookWorkshop-June02/index.html

attrition and reassignment to other projects, which creates some imbalance in the resource for particular areas.

The next five years will see competing changes that NERSC must address. The HPC community will rely much more on open-source and/or semi-supported software, including Linux, archive and global file systems, and complex middleware (e.g., Grid, LDAP authentication, etc.). The SciDAC initiative is creating parallel expertise in highly parallel computing at unprecedented scale. The ISIC projects are in many ways providing an algorithmic support infrastructure that overlaps with discipline-specific scientific support functions that NERSC has provided in the past, including mathematical libraries, profiling and performance tools, as well as community application codes, visualization tools, and even new programming languages. Further, new independent funding for computer architecture investigations is under way, in part due to the High-End Computing Revitalization efforts. These efforts augment NERSC's science-driven system architecture effort, with NERSC concentrating on the applied work not being done in the research community. Finally, increased security concerns and oversight require increased resources, as do additional reporting requirements.

7.1 STAFFING

NERSC will shift staff resources and skills in order to meet the challenges of the next five years: to support the focus on analytics, to support the increased number of systems, and to support increasingly complex science teams. NERSC values its staff and will accomplish the redeployment for the most part by retraining existing staff. If retraining is not feasible, the redeployment will be accomplished by hiring that takes advantage of available openings. The schedule for the shift in staffing will be gradually completed over the next three years.

Staffing changes will shift resources and expertise to supporting critical open-source and infrastructure software, providing support for analytics and data management, and continuing NERSC's science-driven system architecture effort. Since NERSC will maintain the reduced level of staff (10% less than initial 2004 levels), increasing resources in the areas identified above requires shifting resources away from discipline-specific support and from other advanced technology investigations and development. In addition, there will be small shifts in emphasis in system management and security.

7.2 BUDGET

The budget plan is for a flat budget of $38M per year, which is compatible with DOE Office of Science plans. Before 2005 NERSC's budget average was between $28–29M per year, with some additional investment above the original plan such as $3.5M for ~2 TB additional memory on NERSC-3.

The investment strategy for 2006–2010 is very consistent with the past five years, as shown in Table 4.

Table 4
Investment Strategy

Computational systems	~35%
Staff costs	~24%
Infrastructure	~16%
NERSC balance investments	~7%
Overhead/Lab costs	~18%

The infrastructure costs increase approximately 3%, mostly due to ongoing facility electrical costs increasing from ~4% to 12%. This is primarily due to higher usage, not cost per watt, as shown in Figure 18. The increased electrical and maintenance costs are offset by decreases in other infrastructure spending in computational systems, capping media costs, etc.

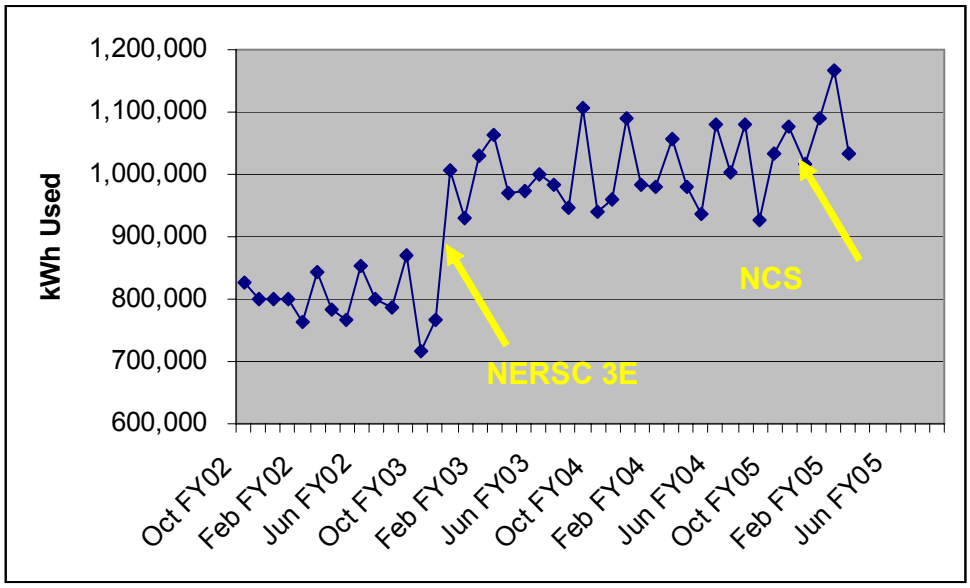


Figure 18. NERSC Center electrical usage, October 2002 to June 2005.

8. Milestones

2005

- NCS enters full service.
  - Focus is on modestly parallel and capacity computing.
  - >15–20% of Seaborg
- WAN upgrade to 10 Gb/s
- Upgrade HPSS to 16 PB. Storage upgrade to support 10 GB/s for higher density and increased bandwidth.
- Quadruple the size of the visualization/post-processing server.

2006

- NERSC-5: initial delivery with possibly a phasing of delivery.
  - 3 to 4 times Seaborg in delivered performance
  - Used for entire workload and has to be balanced
- NCSb enters full service.
  - Focus is on modestly parallel and capacity computing
  - >30–40% of Seaborg
- Replace the security infrastructure for HPSS and add native Grid capability to HPSS
- Storage and Facility-Wide File System upgrade.

2007

- NERSC-5 enters full service.
- Storage and Facility-Wide File System upgrade.
- Double the size of the visualization/post processing server.

2008

- NCSc replaces NERSC-3/3E and NCS.
  - Uses maintenance money for N3E to replace NERSC-3 and NCS with a new capacity system that will be approximately the computational power of the combined systems.
- Increase archive capacity to 40 PB.
- Facility-Wide File System upgrade.
- Upgrade WAN to 40 Gb/s.

2009

- NERSC-6: initial delivery.

    - 3 to 4 times NERSC-5 in delivered performance
    - Used for entire workload and has to be balanced
- NCSc enters full service.
- Storage and Facility-Wide File System upgrade.
- Double the size of the visualization/post processing server.

2010

- NERSC 6 in full service.
- Upgrade WAN link to 100 Gb/s.
- Storage and Facility-Wide File System upgrade.

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.